UNIVERSIDADE D
COIMBRA

Pedro Matos Pinto Santos Filipe

# A Computational Method to Predict the Combinatory Effect of Drugs in Cancer

Dissertation within the Masters in Pharmaceutical Biotechnology advised by Professor Irina de Sousa Moreira and Professor Luís Pereira de Almeida and presented to the Faculty of Pharmacy of the University of Coimbra.

July 2019

**University of Coimbra**

Faculty of Pharmacy

# A Computational Method to Predict the Combinatory Effect of Drugs in Cancer

Dissertation submitted to the Faculty of Pharmacy of the University of Coimbra for the attainment of the degree of Master of Science in Pharmaceutical Biotechnology

*Author:*

Pedro Matos Pinto Santos Filipe

*Advisors:*

Professor Irina S. Moreira, PhD

Professor Luís P. Almeida, PharmD, PhD


FACULDADE DE FARMÁCIA
UNIVERSIDADE D
COIMBRA

July, 2019

> Part of the inhumanity of the computer is that, once it is competently programmed and working smoothly, it is completely honest.

*Isaac Asimov in*
Change! Seventy-One Glimpses of the Future

# Agradecimentos

À Professora Doutora Irina de Sousa Moreira, pela orientação científica do trabalho conducente à elaboração desta dissertação de mestrado. Pela confiança em mim depositada para o desenvolvimento deste projeto. Pelo entusiasmo e disponibilidade demonstrados ao longo deste percurso. Pelo seu exemplo de determinação, otimismo e sucesso que para mim permanecerão como orientação na ciência e na vida.

Ao Professor Doutor Luís Fernando Morgado Pereira de Almeida, meu coorientador interno da Faculdade de Farmácia da Universidade de Coimbra, pela disponibilidade em me guiar neste percurso de forma tão direta, informal e sincera.

Aos investigadores do *Data Driven Molecular Design*. À Joana Mourão pela ajuda e acompanhamento próximos, encorajamento, amizade e partilha; pelo exemplo de integridade e experiência científica e pessoal; pelas sugestões e aconselhamento pertinentes no desenrolar do trabalho; e pela revisão integral desta dissertação. Ao António Gomes pelo papel preponderante que teve, e tem, no meu desenvolvimento pessoal, académico e profissional. Pelo apoio e acompanhamento na elaboração deste trabalho, pelas sugestões pertinentes, e pela revisão desta dissertação. Finalmente, por ser um amigo e um companheiro entre os melhores que Coimbra me deu. À Salete Baptista pelo acompanhamento e amizade, pela perspetiva que aprimorou este trabalho, e pela revisão deste documento. Ao Carlos Barreto pela boa disposição e partilha que muito adicionaram à minha experiência no grupo.

À minha família, em particular aos meus pais e padrasto por todo o apoio, incentivo e aconselhamento ao longo do período em que desenvolvi este trabalho e durante a minha restante estadia na comunidade académica; e pelo encorajamento em todos os projetos que abracei até hoje.

Aos meus amigos, aos de Coimbra e aos de casa, por sempre me apoiarem e servirem de porto de abrigo. Peço-lhes desculpa por não estar tão presente quanto os acostumei, reconhecendo, no entanto o suporte, que mesmo à distância, me deram durante este ano. Um agradecimento especial à Ana, uma excelente namorada, mas também uma excelente amiga a quem agradeço por tudo.

*Muito Obrigado*

# Abstract

Cancer is one of the more significant causes of mortality worldwide. The impact of environmental factors and the growing adoption of risk behaviours have contributed to the high incidence of cancer, especially in low-income countries, where access to screening, early diagnosis, treatment or palliative care is limited. Furthermore, the high tumour heterogeneity and biological complexity make drug-resistance an overwhelming problem. Overcoming this problem can be achieved by finding new and low-cost therapeutic alternatives that do not involve taking the extremely pricy path of original drug R&D. The development of new combinatory therapies is a commonly used cheaper solution that requires fewer clinical experiments when compared to the development process of new drugs, and that often increases the efficacy and reduces the probability of drug resistance. In the last years, the rises of new high-throughput OMICs technologies have made possible the acquisition of large amounts of OMICs data, allowing the unprecedented characterization of cancer biology and behaviour. Dealing with these massive amounts of data so that prolific biological interpretations can be extracted that may aid in the development of more targeted therapies has been a daunting task. Machine Learning (ML) methods are increasingly popular cheaper and faster approaches used to analyse OMICs and integrate this knowledge with other cancer-related data.

In this work, we propose a new ML model for the prediction of an innovative combinatory therapeutic solution, developed using OMICs data that characterize cancer cell lines (specifically expression, methylation and copy number variation) and structural and physico-chemical properties of drugs approved by the FDA for cancer chemotherapy. The best performing approach, an ensemble model comprising a Deep Neural Network, a Random Forest and a Support Vector Machine, achieved 0.74 accuracy, 0.75 precision and 0.90 recall and was suited for the prediction of new combinations for chemotherapy by reliably performing drug screening assays and eliminating less advantageous candidates. We went further ahead and developed also a new database of Membrane Proteins, the most common targets for chemotherapy, and analyse their main interfacial features and characteristics. This database is indeed an important tool for future studies regarding the subject of this work.

**Keywords:** Cancer, Combination Therapy, OMIC sciences, Chemotherapy, Machine Learning.

# Resumo

O cancro é uma das causas mais significativas de mortalidade em todo o mundo. O impacto de fatores ambientais e a crescente adoção de comportamentos de risco têm contribuído para a alta incidência de cancro, especialmente em países subdesenvolvidos, onde o acesso ao diagnóstico precoce, ao tratamento ou aos cuidados paliativos é limitado. Além disso, a elevada heterogeneidade e a complexidade biológica das células tumorais tornam a resistência à terapia um problema emergente. Uma das alternativas para combater o problema das resistências é encontrar estratégias terapêuticas alternativas de baixo custo que não envolvam todas as etapas do processo complexo e oneroso da investigação e desenvolvimento de novos fármacos. O desenvolvimento de novas terapias combinatórias, uma abordagem que aumenta a eficácia e reduz a probabilidade de resistência aos fármacos, é uma solução frequentemente utilizada e relativamente acessível, que requer menos ensaios clínicos em comparação com o desenvolvimento de novos medicamentos. A ascensão de tecnologias Ómicas de elevado rendimento possibilitaram a aquisição de grandes quantidades de dados, desde a génomica, transcriptómica, proteómica até à metabolómica, que têm permitindo caracterizar as células tumorais do ponto de vista biológico e funcional. No entanto, lidar com esta enorme quantidade de dados, de modo a que possam ser extraídos conhecimentos biológicas proveitosos que possam ajudar no desenvolvimento de terapias mais direcionadas, é uma tarefa complexa. Os métodos de Aprendizagem Computacional (AC) são abordagens cada vez mais populares e baratas, para a análise de dados de Ómicas e na integração desse conhecimento com outros dados relacionados com o cancro.

Neste trabalho, propomos um novo modelo de AC para a previsão de novas soluções de terapias combinatórias, recorrendo para isso a dados de Ómicas que caracterizam as linhas celulares de tumores (especificamente para dados de expressão, metilação e variação do número de cópias) e às propriedades físico-químicas e estruturais de fármacos aprovados pela *FDA* para quimioterapia no cancro. A abordagem com melhor desempenho, um *Ensemble Model* composto por uma *Deep Neural Network*, uma *Random Forest* e uma *Support Vector Machine*, obteve uma *accuracy* de 0.74, *precision* de 0.75 e *recall* de 0.90. Este modelo possibilitou a previsão de novas combinações terapêuticas através da realização de ensaios de *screening* de fármacos conducentes à eliminação de candidatos menos vantajosos. Além disso fomos mais ambiciosos e desenvolvemos uma nova base de dados de Proteínas Membranares (os alvos mais comuns para a quimioterapia) que contém as suas principais características e das suas interfaces. Esta base de dados é uma ferramenta importante para estudos futuros no âmbito deste trabalho.

**Palavras-chave:** Cancro, Terapia Combinatória, Ciências Ómicas, Quimioterapia, Aprendizagem Computacional.

# Contents

# List of Abbreviations

AI .....Artificial Intelligence
ADME.....Absorption, Distribution, Metabolism, and Excretion
ANN .....Artificial Neural Network
ATC .....Anatomical Therapeutic Chemical
CCLE .....Cancer Cell Line Encyclopaedia
CI.....Combination Index
CNV.....Copy Number Variation
COSMIC .....Catalogue of Somatic Mutations in Cancer
DNA .....Deoxyribonucleic Acid
DNN.....Deep Neural Network
DT .....Decision Trees
EC50 .....Half Maximal Effective Concentration
EMA .....European Medicines Agency
FDA.....United States Food and Drug Administration
GDSC .....Genomics of Drug Sensitivity in Cancer
HSA.....Highest Single Agent
IC50 .....Half Maximum Inhibitory Concentration
ICGC .....International Cancer Genome Consortium
ML.....Machine Learning
MP .....Membrane Protein
NCI-60 .....National Cancer Institute 60
PCA.....Principal Component Analysis
PPIs.....Protein-Protein Interactions
R&D .....Research and Development
RF .....Random Forests
RNA.....Ribonucleic Acid
RSEM.....RNA-Seq by Expectation Maximization
SMILES.....Simplified Molecular Input Line Entry System
SVM.....Support Vector Machines
TCGA.....The Cancer Genome Atlas
ZIP.....Zero Interaction Potency

# Amino-acid Nomenclature

| Amino-acid | One Letter Code | Three Letter Code |
|---|---|---|
| **Alanine** | A | Ala |
| **Cysteine** | C | Cys |
| **Aspartic Acid** | D | Asp |
| **Glutamic Acid** | E | Glu |
| **Phenylalanine** | F | Phe |
| **Glycine** | G | Gly |
| **Histidine** | H | His |
| **Isoleucine** | I | Ile |
| **Lysine** | K | Lys |
| **Leucine** | L | Leu |
| **Methionine** | M | Met |
| **Asparagine** | N | Asn |
| **Proline** | P | Pro |
| **Glutamine** | Q | Gln |
| **Arginine** | R | Arg |
| **Serine** | S | Ser |
| **Threonine** | T | Thr |
| **Valine** | V | Val |
| **Tryptophan** | W | Trp |
| **Tyrosine** | Y | Tyr |

# List of Tables

# List of Figures

# Introduction

The objective of this chapter is to perform an introduction to the context of the problem that led to the development of this research work and presents its main objectives.

## 1.1 Context and Objectives

Cancer is one of the leading causes of mortality worldwide (Bray *et al.*, 2018) and its incidence is growing, as developing countries continue to adopt unhealthy Western lifestyles (Jemal *et al.*, 2010). Furthermore, the increased mortality in those countries' emphases the importance of the development of affordable cancer therapeutics (Bountra, Lee and Lezaun, 2017), as people and countries with less financial resources have no access to cancer screening, early diagnosis, treatment or palliative care. From an industry perspective, developing new drugs and putting them into the market is extremely expensive. Using already developed drugs but with new purposes is a common undertaken posture by the *Pharma Industry* (Ashburn and Thor, 2004). Consequently, the discovery of new combinatory chemotherapeutic strategies is gaining momentum as a reliable solution to overcome cancer drug resistance (Garcia and Odaimi, 2017; Kalemkerian, 2016).

The development of new high-throughput technologies and computational tools backed the emergence of a large volume of OMICs data (Bennett *et al.*, 2005; Siva, 2008) that is currently used to characterize the high biological complexity of cancer cells. However, dealing with this volume of data is only possible through the use of Machine Learning (ML) and data analysis tools. Academia and industry are, nowadays, starting to take advantage of these new tools, like computation- and automation-based pipelines (Schneider, 2017) to try to extract relevant data from these OMICs knowledge. Some studies attempted to integrate this knowledge with other cancer- and drug-related data to predict the combinatory effects of

well-established drugs, although most of these works are still in an early beginning and include very limited data.

This work aimed to build a new ML model that uses cancer multi-omics and structural drug-related data. This model is a new screening tool for affordable cancer treatment solutions that can quickly point out new therapeutic alternatives by accurately predicting the combination effect of drug combinations submitted for analysis.

## 1.2 Structure of the Dissertation

Apart from this introductory chapter, this dissertation is divided into four chapters according to the following:

- Chapter 1: Background, to clarify how this work's importance for the current knowledge and advancement of cancer treatment with an emphasis on multi-omics data and Machine Learning approaches; At the end, we also propose solutions to solve the problem and expected contributions.
- Chapter 2: Materials and Methods with a detailed description of the pipeline of this work, focused on the two main objectives;
- Chapter 3: Results and Discussion to summarize the outcomes drown from this research work. Parallel to results, an overview of the analysis and interpretation is supplemented, making an evaluation that takes into consideration the main objective of this work and considers current literature;
- Chapter 4: Conclusions and Future Work, to describe a summary of the study, briefly examining its impact, limitations and future work.

# Chapter 1. Background

## 1.1 Cancer in Numbers

Non-transmissible diseases are thought to be responsible for most global deaths. Cancer is one of the leading causes of mortality, and the most significant barrier to increasing overall life expectancy worldwide (Bray *et al.*, 2018). World Health Organization (WHO) estimates that, in 2018, cancer caused approximately 10 million deaths (Ferlay *et al.*, 2018). This organization also states that cancer assumes polarized incidence towards developed countries, as exposed in

Figure 1-1 A, which suggests an association of this disease with ageing as well as a strong relationship with risk factors and behaviours usually present in those countries. These factors include tobacco consumption, exposure to occupational carcinogens, diet, obesity, and environmental factors, among others (Toporcov, Wü and Filho, 2018; Vineis and Wild, 2014). On the other hand, in developing countries, data published by the same organization states that 5.7 million people died from cancer, even though incidence is lower in those countries (Jahan, 2016). This fact emphasises the burden of economic power in health, as people with less financial resources have no access to cancer screening, early diagnosis, treatment or palliative care. This tendency is also illustrated in the estimates for 2018, as displayed in

Figure 1-1 B. Moreover, this incidence pattern is also slowly changing due to the adoption of unhealthy Western lifestyles by developing regions, which, in conjunction with environmental and infectious agents, represent a major concern (Jemal *et al.*, 2010). Nevertheless, only 15% of the world's population is included in cancer registries (Bray *et al.*, 2017, 2018), restricting the possibilities of attaining statically accurate data reports (Ginsburg *et al.*, 2012). Under these rates, global policymakers expressed commitment to apply efforts

on controlling and reducing the outcomes of the cancer epidemic (Prager *et al.*, 2018). In 2017, the World Health Assembly approved a resolution known as the "Cancer Resolution" as a mean of promoting access to cancer treatment and care to all, emphasizing the need to invest in cost-effective medicines and therapeutic strategies (*Cancer prevention and control in the context of an integrated approach*, 2017). In line with this integrated strategy, the United States of America announced the "Cancer Moonshot" Initiative, which provides financial support for the research and development of new and accessible therapeutic strategies for cancer (Lowy *et al.*, 2016; Mayer and Nasso, 2017). In Europe, under the Horizon 2020 framework, the European Union offers funding for the development of technologies that embody, among other objectives, advancing current knowledge and therapeutics in cancer, particularly from the development of personalized medicine, using big data and artificial intelligence (AI) (*Horizon 2020 Work Programme 2020 - Health, demographic change and wellbeing*, 2018).

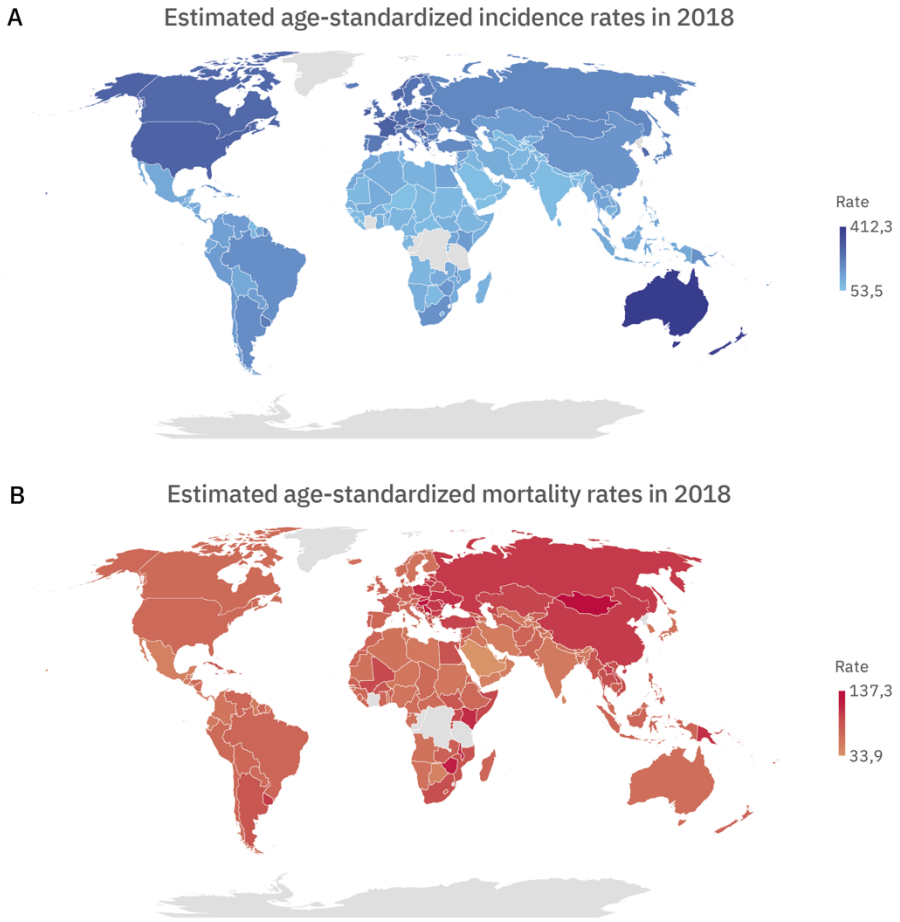



Figure 1-1. Global map representing the estimated mortality (A) and incidence (B) rates of cancer for 2018. The data used to produce these maps was recovered from the International Agency for Research on Cancer (Ferlay *et al.*, 2018).

Early and accurate detection of cancer is a crucial element for a better prognosis and treatment success (Chen *et al.*, 2006; Pellino *et al.*, 2018; Tang *et al.*, 2017). The world has seen considerable improvements in the understanding of the underlying biology and treatment of

cancer over the years, as new imaging (Fan *et al.*, 2017) and molecular techniques are rapidly developed and improved. New biomarkers were also discovered, and its usage was approved for regulatory purposes (Elbehery and Azzazy, 2014; Han, Wang and Sun, 2017; Tang *et al.*, 2017).

## 1.2 Biology of Cancer

Cancer is a complex group of diseases characterized by the presence of abnormal cells with uncontrolled growth that can invade nearby tissues. This factor is mainly caused by genetic changes that often include mutations, such as insertions and deletions in DNA; changes in expression rates (due to changes in sequence of promoters and enhancers); copy number variations (CNV); changes in methylation; and chromosomal translocations (Harris and McCormick, 2010). The sum of these genetic, as well as other non-genetic alterations, further explain the notable tumour heterogeneity and plasticity (Hanahan and Weinberg, 2000) that, associated with the tumour microenvironment, can significantly impact the cell phenotypic behaviour (Hanahan and Weinberg, 2011). Even so, the exact molecular mechanisms underlying cancer initiation, growth, metastasis and resistance to therapy are still poorly understood.

With the development and improvement of high-throughput technologies and computational tools, as well as the downfall of prices of both kinds of instruments (Bennett *et al.*, 2005; Siva, 2008) a large volume of cancer OMICs[1] data is emerging. In fact, genomic data is estimated to generate from 2 to 40 Exabytes ($1.1 \times 10^{12}$ Megabytes) of information per year by 2025 (Stephens *et al.*, 2015), which makes it the most significant domain of Big Data in the near future (competing against data produced by sciences like Astronomy, or big social networks like Twitter and YouTube). New data produced by technologies like mass-spectrometry (e.g. deep -proteomics and -metabolomics), next-generation sequencing (e.g. whole exome sequencing, whole genome sequencing, chromatin immunoprecipitation sequencing, RNA-sequencing, bisulfite sequencing) or microarray enabled the generation of biological data at different levels (e.g. genomic, transcriptomic, proteomic and more recently epigenomic and metabolomics) (Bozic *et al.*, 2013). These genetic profiles can accurately identify and characterize different tumours and cell lines (Gandara *et al.*, 2015), further

[1] OMICs data refers to a field of study in biological sciences that uses high-throughput technologies able to explore the role, relationships and actions of various types of cellular molecules, such as the genes (genomics), mRNA (transcriptomics) proteins (proteomics) or metabolites (metabolomics) that make up the cells of an organism (Debnath *et al.*, 2010).

demonstrating the huge complexity and diversity of human malignancies (Balmain, Gray and Ponder, 2003).

Nowadays, the integration of multi-OMICs approaches is a powerful weapon able to dissect the complex cancer biological mechanisms and to uncover the molecular signatures concerning cellular phenotypes. The first project of high-throughput cancer cell line screening was the National Cancer Institute 60 project (NCI-60) (Shoemaker, 2006), which started in 1984 and during approximately 20 years, screened 60 cancer cell lines against small molecules to identify novel anticancer compounds (Sharma, Haber and Settleman, 2010). The XXI century saw the birth of projects focused on extending the knowledge brought by NCI-60, screening even more cells lines and drugs with the help of new emerging techniques (e.g. the Genomics of Drug Sensitivity in Cancer (GDSC) (Yang *et al.*, 2012); and the Cancer Cell Line Encyclopedia (CCLE) (Barretina *et al.*, 2012)). On the other hand, large international projects, such as the International Cancer Genome Consortium (ICGC) (Zhang *et al.*, 2011) or The Cancer Genome Atlas (TCGA) (Akbani *et al.*, 2014), started accommodating patient tumour multi-OMICs information at an unprecedented scale (specifically genomic, methylomic, transcriptomic and proteomic data). Moreover, smaller projects with other purposes, such as the Catalogue of Somatic Mutations in Cancer (COSMIC), began to store and display somatic mutations information and related details from dispersed oncological databases and publications. The integration and analysis of these data can be performed using manual statistical techniques, although it is a very costly and time-consuming process. Consequently, the use of other approaches (e.g. ML) capable of dealing with this high volume of available data are of utmost importance and will be further discussed in section 1.4.

## 1.3 Cancer Treatment and Drug Resistance

Drug approvals in oncology by the American regulatory agency, Food and Drug Administration (FDA), boosted in recent years. In 2017, 18 new chemical entities reached the market, including utterly new molecules such as copanlisib, a small PI3K inhibitor (ALIQOPA™ – Bayer AG) for relapsed follicular lymphoma. Furthermore, 13 new uses of existing cancer therapies were also approved (Heymach *et al.*, 2018), such as nivolumab, an anti-PD1 monoclonal antibody (OPDIVO® – Bristol-Myers Squibb Co.), for new stages of melanoma; and hepatocellular, colorectal and urothelial carcinomas. This trend was also followed by EMA during the same period, as displayed in Figure 2-2 A. Developing new drugs and putting them into the market is extremely expensive, perhaps explaining the lack of interest from the pharmaceutical industry in new drug R&D (DiMasi, Grabowski and Hansen, 2016). The

number of drugs arriving to shelfs compared to the investment made on research and development processes has lowered significantly (Scannell *et al.*, 2012), as expressed in Figure 1-2 B. Nonetheless, the demand for affordable new drugs continues growing (Bountra, Lee and Lezaun, 2017).

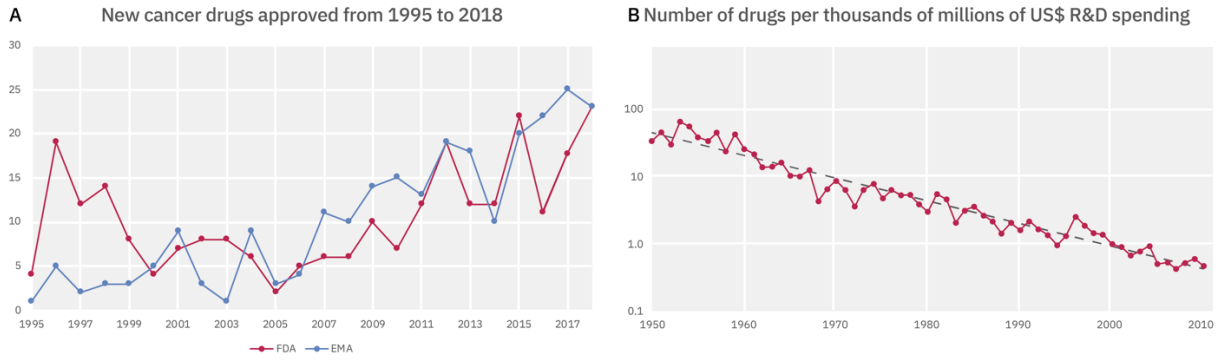



Figure 1-2: A) New therapeutic molecules targeting cancer approved by FDA and EMA between 1995 and 2018. Data corresponding to FDA approvals were retrieved from CenterWatch (*FDA Approved Drugs in Oncology*, 2019), data corresponding to EMA approvals were retrieved from the European Union Open Data Portal (*Medicine data: European public assessment reports (EPAR) for veterinary medicines*, 2018). B) The number of new molecules approved by the FDA per billion US dollars (inflation-adjusted) spent on research and development between 1950 and 2009 (Adapted from Scannet et al., 2012).

The drug development process is not only costly, but also very time consuming, reducing the time on which a drug is under industrial protection (Abrantes-Metz, Adams and Metz, 2004; DiMasi, 2001), and therefore decreasing the profit from launching it to the market (see Figure 1-3) (Sinha and Vohora, 2018). As means of withstanding these factors, pharmaceutical companies are turning themselves to alternative ways of proposing new medicines and therapeutics, that include the focus on precision medicine (Dugger, Platt and Goldstein, 2017), open-source R&D (Munos, 2006), drug repositioning (Ashburn and Thor, 2004), production of biologics (Munos, 2009), and the use of computation- and automation-based pipelines to analyse cancer-associated data (Schneider, 2017).

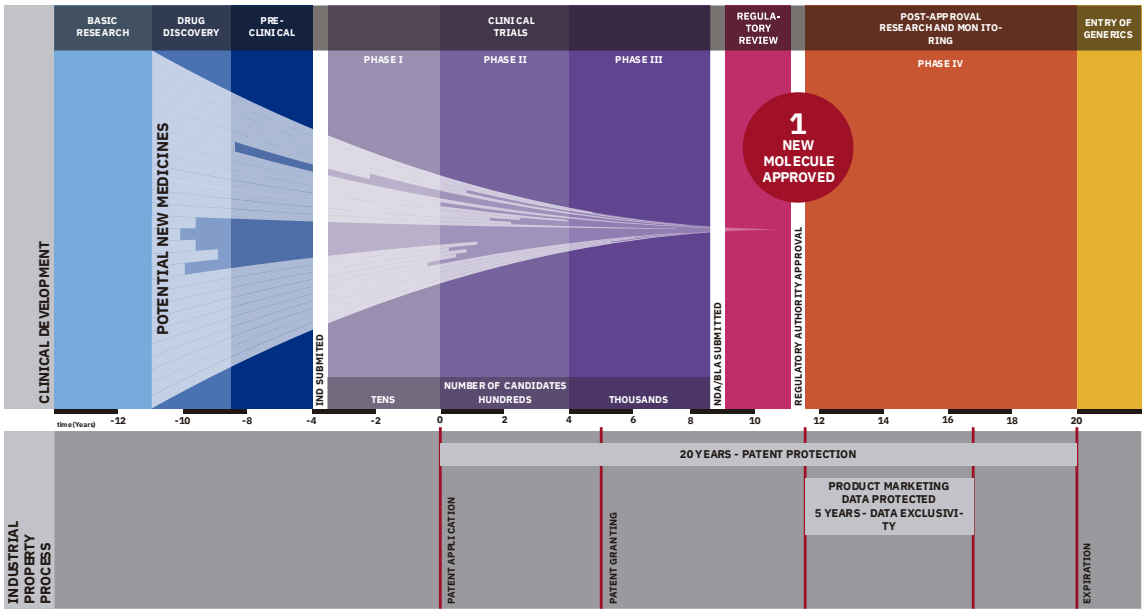



Figure 1-3: The drug development process and the industrial property granting procedure.

Cancer treatment depends largely on the type and stage of the tumour. Among the different available treatments (surgery, radiation, stem cell transplant, chemo-, immuno-, and hormone- therapy), surgery continues to be the primary treatment for most solid tumours, a process that, besides being very invasive, cannot eradicate the disease entirely due to poor cellular differentiation in most tumours (Chabner and Roberts, 2005). In the past, most conventional therapies treated cancer as a homogeneous disease with several drugs targeting random DNA, inducing a non-specific DNA-damaging cytotoxic mechanism. However, these drugs were not usually targeted to cancer cells, which, in higher concentrations, may result in serious side effects to the patient (Kummar *et al.*, 2006; Lotfi-Jam *et al.*, 2008). Currently, mechanisms of action of anticancer drugs are based upon entirely different approaches (Patel *et al.*, 2014). Most of these drugs act on well-defined targets or biological pathways based on the molecular and cellular characteristics of cancer cells, therefore improving cancer treatability. More than half of all known drugs on the market target membrane proteins (MPs) or in some cases, their ligands. In cancer, MPs are highly relevant since they usually suffer modifications (e.g. changes in protein composition of membranes, protein expression or glycosylation) during tumorigenesis, making them suitable biomarkers either for diagnosis or therapy (Grimm *et al.*, 2011; Kampen, 2011). However, one of the major drawbacks for the use of MPs as biomarkers is that the structure and function of most of them are still poorly understood, with less than 4% of the crystal structures present in Protein Data Bank (statistics from July 1$^{st}$, 2019) (Berman, Henrick and Nakamura, 2003). Therefore, the development of new tools and the determination of more structures that can further characterize MPs is of utmost importance since it will allow exploiting new cancer therapies.

Although conventional chemotherapy and targeted antineoplastic drugs are still used and effective in most of the cases, the rise of drug resistance is currently an overwhelming problem (Chen and Malhotra, 2015; Igney and Krammer, 2002). Cancer resistance to chemotherapeutics can be divided in two general categories: intrinsic or acquired. Intrinsic resistance occurs before any exposure to treatment and can be present in roughly 50% of all cancer patients (Verheul *et al.*, 1998; Zahreddine and Borden, 2013). These mechanisms can include resistance-mediating factors that pre-exist in tumour cells, or host factors such as poor absorption and rapid metabolism, that often lead to the reduction of total concentration of the drug in the gastrointestinal tract and bloodstream (Foo and Michor, 2014). Acquired resistance, on the other hand, occur in sensitive cells following the administration of chemotherapy, being mainly caused by mutations in cancer cells and other adaptive responses such as increased expression of the therapeutic target and activation of compensatory signalling pathways (Davis, Chen and Shin, 2008). Furthermore, other chemo-resistance factors can also contribute to chemoresistance such as intra-tumour heterogeneity (by positive selection of a minor drug-resistant tumour subpopulation), tumour microenvironment, epigenetic changes and cross-talk[2] (Hu *et al.*, 2016). These resistance mechanisms support the urgent need of new or alternative approaches for cancer treatment, including the association of two or more chemotherapeutic agents (discussed in section 2.3.1).

### 1.3.1 Combinatory Therapeutics

The pharmacological concerns involving efficacy and safety of monochemotherapy associated with the tumour heterogeneity, microenvironment and interconnection of multiple disease pathways make this an inadequate and insufficient approach to treat cancer (Ibrahim *et al.*, 2012; Miao and Huang, 2015). Alternatively, combination chemotherapy, or polychemotherapy (simultaneous administration of two or more chemotherapeutic drugs), is a common treatment to avoid some of the disadvantages of monochemotherapy and overcome cancer drug resistance (Garcia and Odaimi, 2017; Kalemkerian, 2016). In opposition to monochemotherapy, combination therapy can modulate different targets and signalling pathways simultaneously (Hu *et al.*, 2016), mainly due to the diverse physico-chemical and structural characteristics of each drug (Figure 1-4). These polychemotherapeutic strategies follow a set of defined rules, including nonoverlapping toxicity of drugs in combination, non-cross resistance and mandatory enhancement of cell kill efficacy (Mayer and

[2] Cross-talk refers to cross regulation of two biological entities (Dogterom and Koenderink, 2019).

Janoff, 2007). Additionally, the positive effect of polytherapy can be further enhanced by modulating the delivery system that conveys this cocktail of drugs by, for example, designing nanoparticle delivery systems that improve stability and circulation time in the bloodstream; targeting specific tissues or cells; or improving intracytoplasmic delivery (Simões *et al.*, 2004).

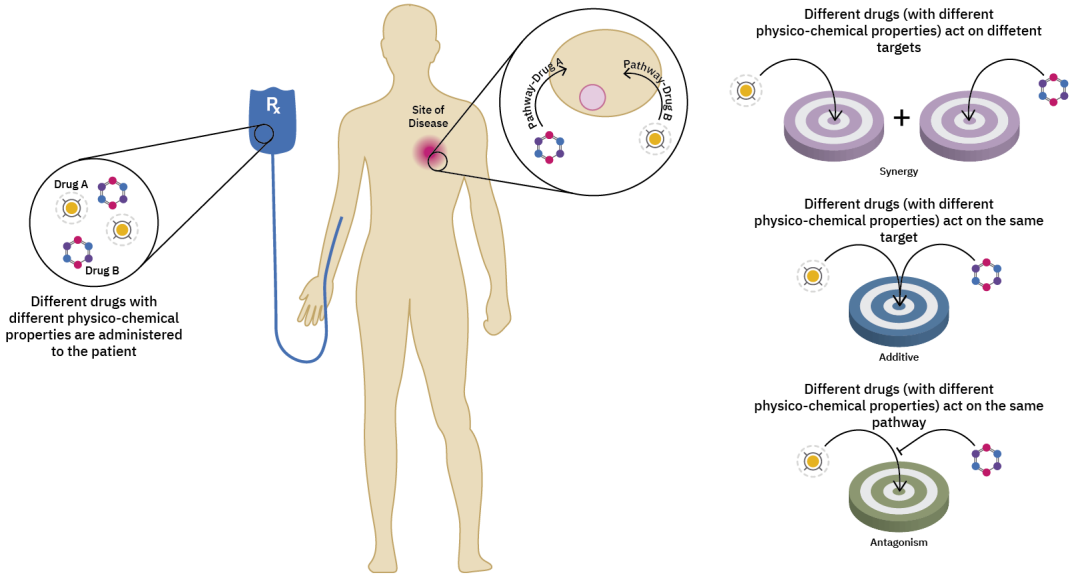



Figure 1-4: Schematic representation of the mechanisms underlying combinatory chemotherapy.

Combination therapies are also required to be approved by regulatory agencies, similarly to mono-therapeutic strategies (Webster, 2016). Supplementary Table 1 lists currently approved combinatory chemotherapeutic strategies for most types of cancer. According to the Loewe additivity model (Loewe and Muischnek, 1926) polychemotherapy can be separated into three groups depending on the therapeutic effects generated, and the targets and pathways involved: synergistic, additive and antagonistic. In synergistic effects, the administration of drugs in combination is greater than the summed effects of the drugs in monotherapy, and the targets or pathways tackled by these drugs must be different. Additive effects occur when the administration of drugs in combination is greater than or equal to the summed effects of the drugs in monotherapy, and the targets or pathways involved may be the same. Examples of these kinds of effects are portrayed in Table 1 A drug combination is classified as antagonist if the presence of one molecule interferes in the mechanism of action of others in the combination, cancelling its therapeutic effect mostly by binding to the target without producing therapeutic response (Kenakin, 2012).

Table 1-1: Examples of synergistic and additive effects of drug combinations applied to cancer.

| Drug A | Drug B | Combinatory Mechanism |
|---|---|---|
| *Synergistic Effects* | | |
| Oxaliplatin<br>• DNA adduct - binds GG, AG and TACT sites (Chaney *et al.*, 2004);<br>• Causes DNA strand break and non-DNA initiated apoptosis (Woods and Turchi, 2013);<br>• Effect of oxaliplatin's DNA adduct may be partially reduced by certain mutant DNA TOP1 acting on DNA adduct to generate different topoisomers (Kobayashi *et al.*, 1993). | Irotecan<br>• DNA TOP1 inhibitor (Koizumi *et al.*, 2004). | • Irotecan acts as an adjuvant of oxaliplatin by inhibiting DNA TOP1 partially offsetting the counteractive activity of mutant enzymes (Koizumi *et al.*, 2004). |
| *Additive Effects* | | |
| Doxorubicin<br>• DNA intercalator – binds AT regions (Kellogg, Scarsdale and Fornari, 1998). | Trabectedin<br>• Interacts with DNA repair system - forms covalent guanine adducts in DNA minor groove (Zewail-Foote *et al.*, 2001). | • Both affect DNA in a non-interfering way, expanding the therapeutic effect (Pautier *et al.*, 2015). |

Through the years, multiple methods were proposed to classify the combinatory chemotherapeutic effect apart from the Loewe additivity method (Loewe and Muischnek, 1926). These newer methods include median-effect (Chou-Talalay Method) (Chou, 2010), Bliss independence (Bliss, 1939), Highest Single-Agent approaches (HSA) (Foucquier and Guedj, 2015) and Zero Interaction Potency (ZIP) (Yadav *et al.*, 2015). Through the referred methods it is possible to calculate Combination Index (CI) and classify the combination effect as additive, synergistic or antagonistic (Chou and Talalay, 1984), following the concept designed by Loewe and colleagues in 1926.

Median-effect and Loewe additivity approaches calculate CI (Equation 1-1) comparing the administered doses and the IC50 or EC50[3] of each drug, represented as $p$. These

[3] Half Maximal Inhibitory Concentration (IC50) and Half Maximal Effective Concentration (EC50) indicate how much of a particular drug or other substance is needed (measure of the potency) to inhibit a specific biologic process (or component of a process). IC50 is the concentration of a drug/substance required to produce 50% inhibition *in vitro* (Brody, 2018). EC50 is the concentration of an agonist that produces 50% of the maximal possible effect of that agonist (Neubig, 2003).

approaches bear in mind the *dose equivalence principle* (for a given effect, dose *a* of drug A is equivalent to dose $b_a$ of drug B, and reciprocally) and the *sham combination principle* ($b_a$ can be added to any other dose b of drug B to give the additive effect of the combination). In the Loewe additivity method, the additive effect of drugs A and B depends on the individual dose-effect curves of each drug (Equation 1-2), also assuming that the drugs have a *constant potency ratio*, R ($R = \frac{A}{B}$). Experimentally, dose-effect curves with constant potency ratio (Figure 1-5 A) have a ratio of doses at every level of effect and are parallel on a log-dose scale, having equal individual maximum effects (Foucquier and Guedj, 2015; Tallarida, 2012). From this, it is possible to define a relationship between the dose pairs from the combination cocktail and each individual drug dose (Equation 1-3). The types of combinatory effects can be drawn by calculating CI (Equation 1-4 and 1-5) and by representing the data in an isobologram (Figure 1-5 B). In addition to the method proposed by Loewe, the median-effect approach considers two more components: the ratio of cancer cells affected by the drugs and a new constant bearing the influence of sigmoidicity of the dose-effect curve.

$$CI = \frac{D_A}{p_A} + \frac{D_B}{p_B} \qquad \text{Equation 1-1}$$

where $D_A$ and $D_B$ are the doses of drugs A and B; and $p_A$ and $p_B$ are the IC50/EC50 of each drug respectively

$$Effect(a + b) = E_A(a + a_b) = E_B(b_a + b) = E_{AB} \qquad \text{Equation 1-2}$$

where $E_A$, $E_B$ and $E_{AB}$ are the measured effects on the dose-effect curve of drug A, B and combined respectively

$$a + a_b = A \iff a + b \times R = A \iff a + b \times \frac{A}{B} = A \qquad \text{Equation 1-3}$$

$$\frac{a}{A} + \frac{b}{B} = 1 \qquad \text{Equation 1-4}$$

$$\frac{a}{A} + \frac{b}{B} = CI \qquad \text{Equation 1-5}$$

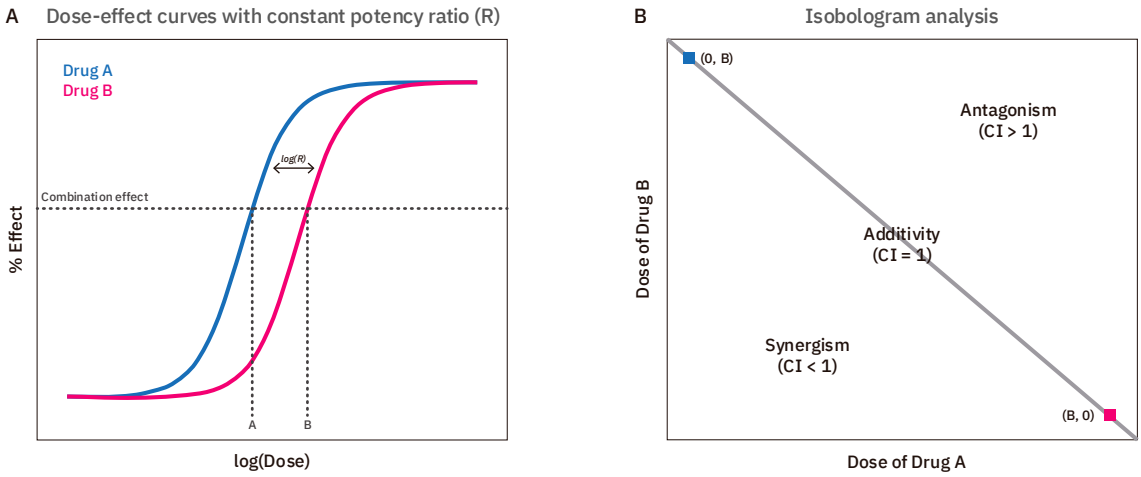



Figure 1-5: Dose-effect curves and isobologram analysis. A) represents the dose-effect curves of drugs A and B separated by the logarithm of the constant R. B) represents an isobologram analysis of the combination of drugs A and B, from which the combinatory effect can be drawn calculating the CI.

The other mentioned methods (Bliss independence and Highest single agent) follows a much reductive way. Bliss independence allows the calculation of CI by comparing the effect of the single drug in the given doses with the observed inhibitory effect of the same dose combination (Equation 1-6 and 1-7). On the other hand, by using HSA, CI is obtained comparing the maximum effect of the single drugs with the inhibitory effect of the combination (Equation 1-8). At last, the ZIP model takes advantage of both Loewe and Bliss models by comparing the dose-response shift between individual drugs and their combinations.

$$E_{add} = E_a + E_b(1 - E_a) \quad \text{Equation 1-6}$$

$$CI = \frac{E_{add}}{E_{observed}} \quad \text{Equation 1-7}$$

$$CI = \frac{max(E_a, E_b)}{E_{ab}} \quad \text{Equation 1-8}$$

## 1.4 Artificial Intelligence and Machine Learning

### 1.4.1 Use of Machine Learning in Data-Driven Decisions

As presented in the previous sections, newly developed high-throughput OMICs platforms, among other technological advances, made possible the acquisition of a large amount of biological cancer data like never seen before. The integration of different types of complex and high-dimensional drug-omics cancer data (from the cell to drug and phenotype) is of utmost importance for understanding cancer's complex biological system and uncovering key altered pathways, that can further help to improve the patient outcomes by designing

more targeted therapies. Integrating, organizing and interpreting this vast amount of data and knowledge to uncover important biological finding is almost impossible using only manual analysis. Computational techniques, in particular ML, a subset of Artificial Intelligence (AI) has been a common approach to overcome the technical barriers of dealing with big data analysis (Auffray *et al.*, 2016; Ritchie *et al.*, 2015). These methodologies help to bridge the gap to the wet lab, appealing to the scientific community as a less costly and time-effective approach (Zhang *et al.*, 2017).

Concerning AI, its definition has evolved from a human-centred perspective (in a sense in which Machine Intelligence should mimic the Humans' thinking and behaviour) to a more rational viewpoint, on which the machine, mainly applying a combination of mathematics and engineering, can automate a set of activities, such as decision-making, problem-solving and learning (Russell and Norvig, 2010). From a general perspective, the process of learning implies that the agent is improving its performance on tasks as it observes the world (Murray, 2003). This procedure depends on the prior knowledge of the agent, the representation of the knowledge that the agent has access to and the availability of feedback to learn from.

The learning concept is transferred to machines via ML, thus avoiding the need of explicitly programming a machine to perform a specific task. The agent (in this case, the machine) performs calculations on a given set of descriptive data to make a statistical representation of those records by automatically recognizing patterns within it (Bishop, 2006). This representation is compared to the available knowledge on the subject in analysis and, if they match, it generalizes to unknown examples taking into the account the previous learning procedure (Pastur-Romay *et al.*, 2016). A common way of classifying ML types is according to the availability of learning feedback through labelled data as Unsupervised, Supervised and Semi-Supervised, being the first two the most frequently used and further detailed in the next sections (Putin *et al.*, 2018). Furthermore, Reinforcement Learning can be found as another type of ML. However, in this case, the machine learns from a series of rewards or punishments.

The organization of the next sections takes into consideration the design of the algorithm and the learning process of each method with a further explanation of the basic parts and characteristics of each one (Figure 1-6).

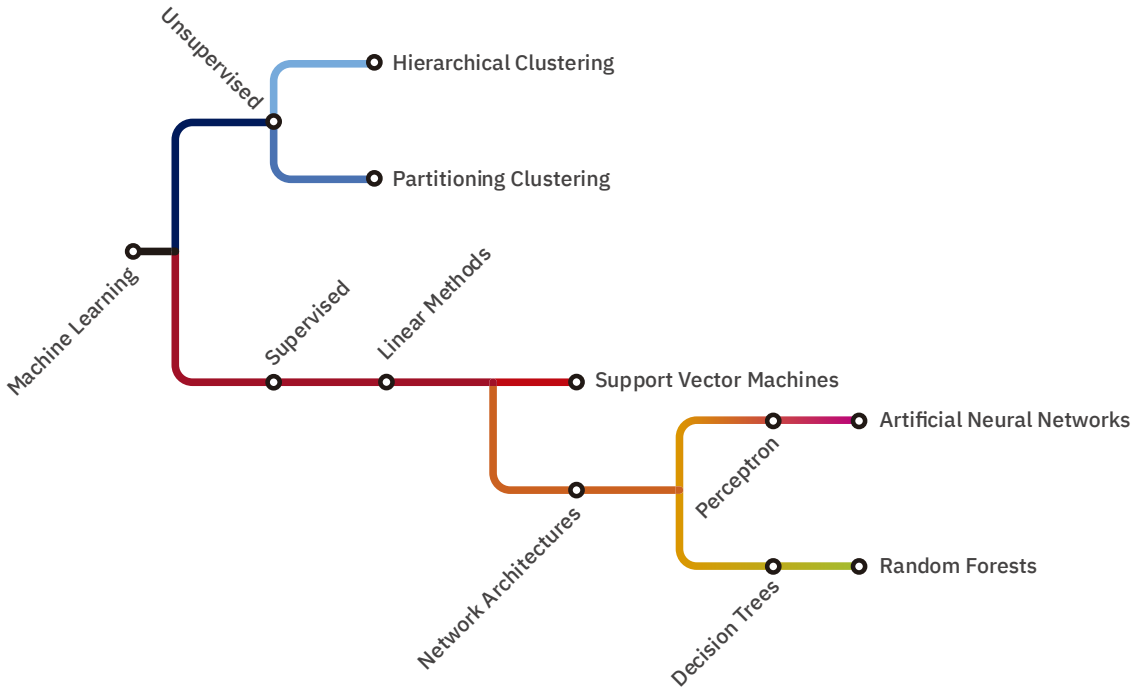



Figure 1-6: Machine Learning types mentioned in this section.

#### *1.4.1.1 Unsupervised Learning*

In Unsupervised Learning the machine draws predictions from datasets by simply observing raw and unlabelled input data (Becker and Plumbley, 1996). Datasets contain samples, also called instances[4], containing descriptors of the samples, commonly known as features[5] (Mohri, Rostamizadeh and Talwalkar, 2012) (e.g. a dataset comprising descriptors of genes, like length of telomers or percentage of CpG islands). The most known Unsupervised Learning method is Clustering (Russell and Norvig, 2010). Nevertheless, other approaches of unsupervised learning include Anomaly Detection and some types of Neural Networks.

Clustering, or cluster analysis, is roughly defined as grouping samples of a given dataset by some criterion of similarity. However, this criterion varies according to the objective of the analysis (Estivill-Castro, 2002). Clustering can be performed using Hierarchical or Partitioning methods (Ferligoj and Batagelj, 1983; Fraley and Raftery, 1998). Hierarchical methods progressively divide the samples into groups arranged in according to similarities between them. This clustering method can be performed either top-down, by progressively grouping different instances, thus forming a large group that accommodates the whole smaller groups (Agglomerative Hierarchical Clustering), or bottom-up, by iteratively splitting the instances into smaller groups, creating new and more specific groups at each step (Divisive

[4] Instance (example, case or record) refers to a single object of the dataset from which a model will be learning or predicting.

[5] Feature (attribute, field or variable) is referred as a descriptive characteristic of an instance.

Hierarchical Clustering). Partitioning methods relocate samples by moving them from one cluster to another. The simplest and most commonly used model for Partitioning clustering is the $k$-means algorithm (Jain, 2010), which often requires a number of $K$ clusters $(C_1, C_2, \ldots, C_K)$ to be set by the user. This kind of clustering procedures iteratively relocates samples between the clusters until a certain error criterion is minimized, classifying them as Error Minimization Algorithms. Ultimately, this error criterion measures the "distance" of each instance to its representative value (a point in the centre of the group called the centroid) (Wagstaff *et al.*, 2001). Commonly, the model employs the Sum Squared Error (Equation 1-9) to determine the error and to split data into the defined clusters (Rokach and Maimon, 2005).

$$SSE = \sum_{i-1}^{n} \left(y_i - f(x_i)\right)^2 , \qquad \text{Equation 1-9}$$

where $y_i$ is the $i^{th}$ representative value and $f(x_i)$ the predicted value of $y_i$.

### *1.4.1.2 Supervised Learning*

In Supervised Learning, the agent learns by receiving feedback from a set of examples of labelled data (e.g. a dataset comprising descriptors of genes and labels that classify each sample with the type of cancer which they belong to). After fitting[6] the training set, usually corresponding to 70% of the whole dataset (since the other 30% remain reserved for the test set - model evaluation), the algorithm can predict the label associated with new and unseen samples using descriptors that characterize it. Supervised Learning algorithms are used to solve problems associated with regression – if labels are numbers contained within a certain continuous domain (e.g. prediction of IC50 of drugs), or classification – if labels are categorical (which also contains binary labels) and finite (e.g. predicting the possible target of different drugs according to its morphological features). Generalizing this concept, for a given set of $N$ examples $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$ where $x_i, y_i$ are descriptors and label, respectively, of the $i^{th}$ sample and $y = f(x)$. The Supervised Learning model $M$ must discover a hypothesis function $h$, where $h(x) \sim f(x)$ (Russell and Norvig, 2010). The accuracy of the final model (based on the best hypothesis) is evaluated by comparing it with descriptors and labels of the test set which comprises new samples from the same dataset used to train it (these instances cannot be used to train the model). If $M$ can successfully predict the labels from this new unknown set, it means that it can generalize (Behzad *et al.*, 2009), thus accomplishing its purpose. However, two major problems can occur if the model is incapable of generalizing: Underfitting or Overfitting. Underfitting occurs when the model is incapable of finding

[6] Fitting is the process of adjusting the hypothesis ($h$) to the data used to train the model.

noteworthy patterns in the training set. On the other hand, overfitting is observed when it is over-trained, thus memorizing the training set without being capable of generalizing to new examples (Aalst, van der *et al.*, 2010; Kouvaris *et al.*, 2015) (Figure 1-7).

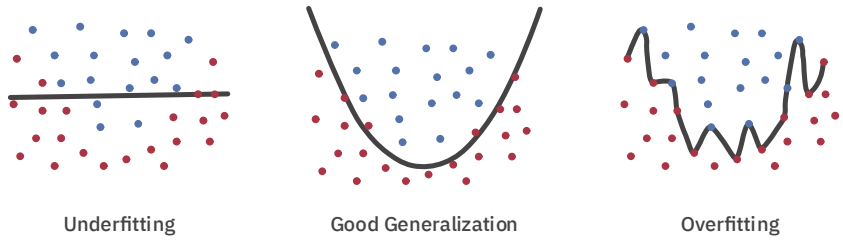



Figure 1-7: Concepts of underfitting, generalization and overfitting.

#### *1.4.1.2.1 Linear Models*

Linear models are simple algorithms that are based on linear relationships between samples' attributes (features) and their labels (classes). These models can make good predictions if data is linearly separable, in a much quicker way than nonlinear classifiers (Yuan, Ho and Lin, 2012), since the relationship between features and classes (Equation 1-10), a vector commonly represented as $\phi(x)$, is monodimensional (Equation 1-11). This is illustrated in Figure 2-8, where a linear function $h$ is separating two regions ($R_1, R_2$) where, in this case, samples corresponding to two target classes must reside. Thus, in practice, samples in $R_1$ (above $h$) belong to class "A" and samples in $R_2$ (below $h$) belong to class "B". A similar approach is taken for multiple classes but, instead, each region is defined by two linear functions (Bishop, 2006). On the other hand, linear models can also be used for regression purposes. However, in this case, the outputs' space is defined by $h(x)$ itself.

$$h(x) = w^T \phi(x) + b,$$
where $w^T$ is a weight vector and $b$ is a bias value

Equation 1-10

$$\phi(x) = x$$

Equation 1-11

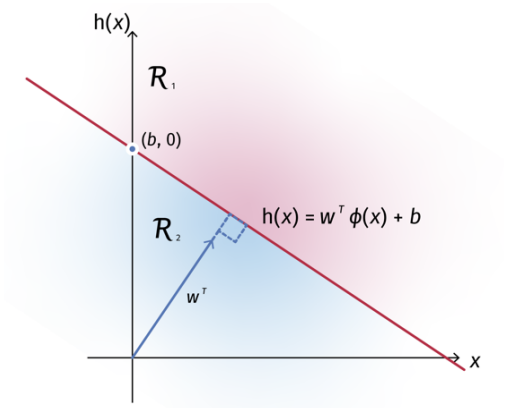



Figure 1-8: Linear model representation. The x-axis represents the $\phi(x)$ vector, the y-axis represents the calculated $h$ value for each $x$.

Examples of linear models include Support Vector Machines (SVM) (Cortes and Vapnik, 1995). Nevertheless, SVMs create representations of multidimensional $\phi(x)$, instead of dealing with monodimensional features' spaces. This method represents each sample $(x_i, y_i)$ in a hyperspace of $n$ dimensions; thus, each data point is defined by the $n$ dimensions of the features within the vector $x_i$ (Equation 1-12) (Ziegel *et al.*, 2007). These data points are placed in specific regions of the dataspace, such that a hyperplane can separate two regions which assume values of ±1 assigned to $y_i$, used to classify the instance as +1 if the sample belongs to the class or −1 if the instance is classified as something else. SVM's designation comes from the calculation of two margins by each side of the hyperplane splitting the two regions. These margins are defined to maximize the distance between the hyperplane and the closest data point, or points if there are more than one (called support vectors) (Figure 1-9). By applying this method, the number of hyperplanes capable of splitting the data points is significantly reduced, giving the algorithm its optimal stability.

$$x_i = (x_0, x_1, x_2, x_4, \ldots, x_n)$$ Equation 1-12

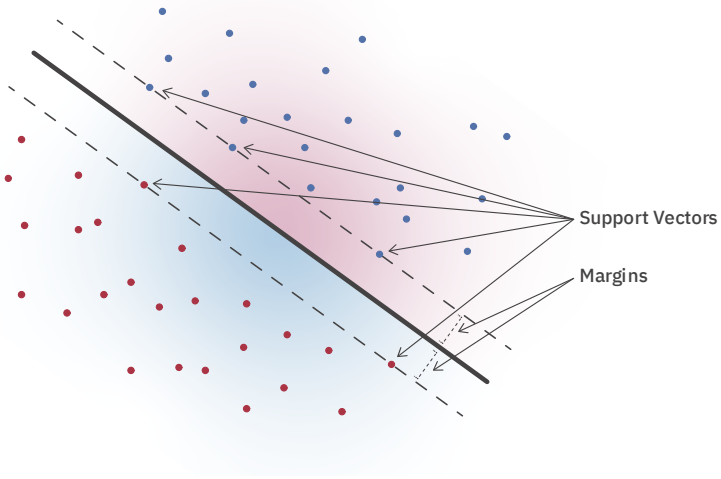



Figure 1-9: Support Vector Machines, in the case of linearly separable data. Highlighted in red and blue are the data points of each region.

#### *1.4.1.2.2 Network Models*

Most currently used network-based ML algorithms are based on Decision Trees (DTs) and Artificial Neural Networks (ANNs). Decision Trees require less parameter tuning and offer an often-quicker ML solution (James *et al.*, 2013). However, they show less robustness and tend to lead to overfitting (Bramer, 2007). These methods predict the label associated with an instance by making it sequentially travel from root nodes ($x_0$) to leaf nodes ($x_n$) (Figure 1-10 – A). In each node, the successor node is chosen by splitting the input space according to predefined values or according to the set of labels most capable of lowering the uncertainty of a given feature (Russell and Norvig, 2010). Random Forests (RF) are classifiers built with a collection of DTs organized in ensemble[7] (Figure 1-10 – B). The training dataset travels across all trees. The final classification is the predicted by the majority of these trees (Shalev-Shwartz and Ben-David, 2014).

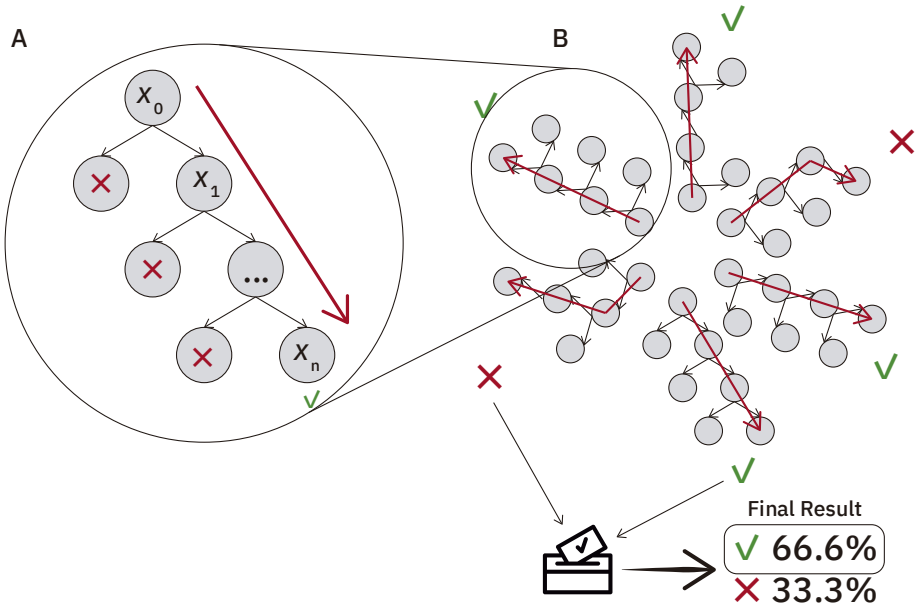



Figure 1-10: Decision Trees (A) and Random Forests (B).

The perceptron, first proposed by Frank Rosenblatt in 1958 (Rosenblatt, 1958), is a type of binary classifier and is the basis of the principle regarding ANN architecture. Perceptron's mimic the multipolar neural cells found in the brain: the inputs resemble the dendrites found in the cell body, the sum and threshold function the axon, and the output the axon terminals of the neurons (Figure 1-11 - A). Conceptually, both perceptron's and ANNs are networks composed of nodes connected by edges (a graph-based model). These edges have

[7] An ensemble of classifiers is a set of classifier models whose individual decisions are combined in some way (typically by weighted or unweighted voting) to classify new examples (Dietterich, 2000; Zhou, 2012).

weights ($w$) that determine the strength of the connection and propagate activation values across the network (Figure 1-11).

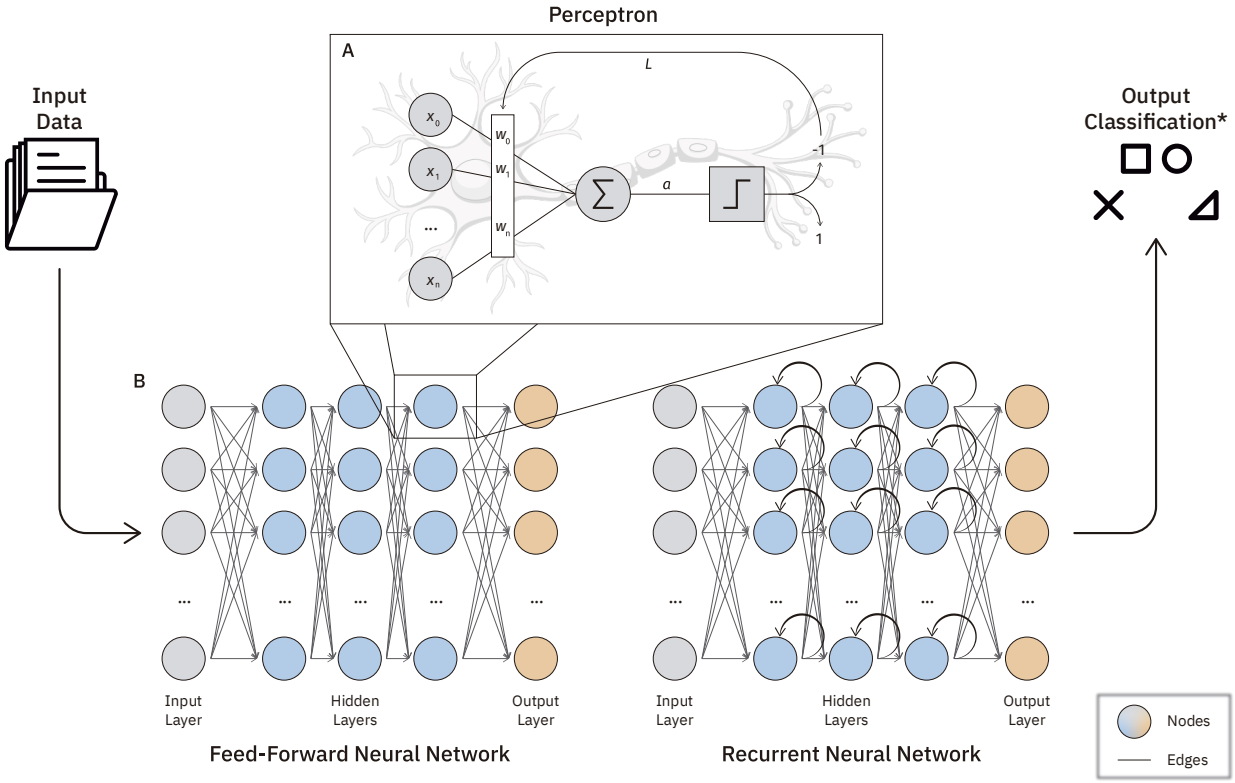



Figure 1-11: Network machine learning models. A) Representation of a perceptron. B) Example of Deep Artificial Neural Network. * Although the only represented objective is classification, ANNs can perform other kinds of tasks. Figure created using resources from Freepick.com.

A supervised perceptron algorithm decides if a particular input belongs to a specific class by analysing a set of features relative to the sample in analysis. These features are multiplied by weights $w$, and the product of all features by its weights is summed, resulting in activation $a$. In the case of the first perceptron, the final output is given by a hard threshold function (Equation 1-13), although nowadays we can find a broader set of activation functions besides hard thresholds (Russell and Norvig, 2010). The output of this function is used to classify the instance as positive, if the sample belongs to the class ($f(a) = +1$), or negative otherwise ($f(a) = -1$). The model learns from changes in $w$, resulting from a loss function ($L$), in a process called backpropagation (Bishop, 2006) (Figure 1-11 - A). Backpropagation works as the model is presented with the same samples' multiple times (learning epochs[8]) by means of adjusting $w$.

$$f(a) = \begin{cases} +1, & a \geq 0 \\ -1, & a < 0 \end{cases}$$ Equation 1-13

[8] One epoch consists of one full training cycle on the training set. Once every sample in the set is presented to the model, it restarts in the beginning of the training dataset, introducing the second epoch.

ANNs can assume a large variety of architectural differences that may arise from the number of nodes composing the network, as well as the layers on which those nodes are contained. Simpler models include single-layer neural networks, which are built only by the input and output layers, a type of architecture on which the earlier mentioned perceptron is comprised. ANN are data-driven models (Schwabacher, 2005), nevertheless, the same model cannot be used in all datasets (Young *et al.*, 2015). Hyperparametrization allows ANNs to be tuned and achieve better prediction results depending to the problem in analysis. For example, models may also contain more layers, namely hidden layers connecting the inputs to the outputs. If an ANN comprehends more than one hidden layer, it is considered a Deep Neural Network (DNN). Neural networks may present more than one target variable. In categorical problems, in which the objective is to classify instances in more than two classes, the number of output neurons corresponds to the number of classes defining the data. Neural Networks may further assume two fundamentally distinct forms of performing error propagation and be classified accordingly, as feed-forward neural networks or recurrent neural networks (Figure 1-11 - B). Feed-forward Neural Networks have its edges establishing connections in only one direction – from the inputs "upstream" to the outputs "downstream", without recording any other internal state, rather than the weights in each edge. In contrast, recurrent neural networks allow the connection of nodes located within the same layer using a technology named long short-term memory (Chen *et al.*, 2018), saving more internal data and functions than feed-forward neural networks (Hochreiter and Schmidhuber, 1997), which make them especially useful for problems involving processes (eg. speech recognition). As in the perceptron, ANNs learn by fitting changes in the weight vectors performed by a chosen loss function (in fact, ANNs are also called multilayer perceptrons). These functions often depend on the comparison between the value predicted by the model and the real class (inputted by the user). This aspect represents a problem in hidden layers as the training data does not mention the output given by these. The solution to this problem is based on the back-propagation of the error calculated in the output layer, thus creating a gradient of the loss function to readjust the weights in each hidden layer (Li *et al.*, 2012). Loss functions and the function used to adjust the weight vectors according to it (called optimizer) are also hyperparameters. Further hyperparameters include learning rate (the rate in which weight vectors are updated) and learning rate decay (the ratio to which learning rate is updated in each epoch).

Selecting the right supervised machine learning model depends on the problem that the user is willing to solve, as there is a wide variety of mathematical structures to deal with

different type and quantity of data, computational complexity, *et cetera* (Jordan and Mitchell, 2015). In the last few years, different supervised ML methods were used in OMICs data to predict the best therapeutic strategy based on the genomic profile of the patient (Kalari *et al.*, 2018), disease-related *loci* (Leal *et al.*, 2019), patient prognosis (Feng *et al.*, 2019; Kim, Oh and Ahn, 2018; Long *et al.*, 2019; Yu *et al.*, 2019), disease biomarkers (Bravo-Merodio *et al.*, 2019) and best treatment combination scores (Gilvary, Dry and Elemento, 2019; Janizek, Celik and Lee, 2018). Although these methods use OMICs information as features for training ML models, in most cases the use of these data is restricted to only one type of information (e.g. genomic, transcriptomic, proteomic). Furthermore, commonly used datasets present a reduced number of samples and a high number of attributes, which may raise concerns regarding the generalization ability of those models. Besides OMICs data, some studies have developed ML methods with data from physico-chemical and structural properties of anticancer agents to predict Absorption, Distribution, Metabolism, and Excretion (ADME) properties, to identify new uses for existing drugs (repurposing) or even to identify new drug candidates (drug discovery). To take full advantage of the vast repertoire of features that may be related with cancer biology and behaviour, we propose as main objective, to build a ML model capable of predicting anticancer drug combinations for chemotherapy, by using not only physico-chemical and structural properties of these agents but also multi-omics features, filling a gap in the use of multi-source data.

# Chapter 2. Materials and Methods

This chapter provides the material and methods used in this study. It is split into two parts: i) the construction of a Chemotherapeutic Combinatory Effect Prediction ML model and ii) the thorough assessment of drug targets, Membrane Protein Dimer Interfaces, structural and genomic-conservation characteristics.

## 2.1 Chemotherapeutic Combinatory Effect Prediction

An in-depth explanation of the methods used in the construction of a Chemotherapeutic Combinatory Effect ML model is addressed herein. In particular, the combinatory drug phenotypic data acquisition from NCI-ALMANAC (Holbeck *et al.*, 2017), the mining of these drug features using the Python package Mordred (Moriwaki *et al.*, 2018) and the OMICs data retrieval from CCLE database (Barretina *et al.*, 2012) (Figure 3-1). The handling and pre-processing of these data and the overall model construction and deployment using Scikit Learn (Pedregosa *et al.*, 2011) is also presented.

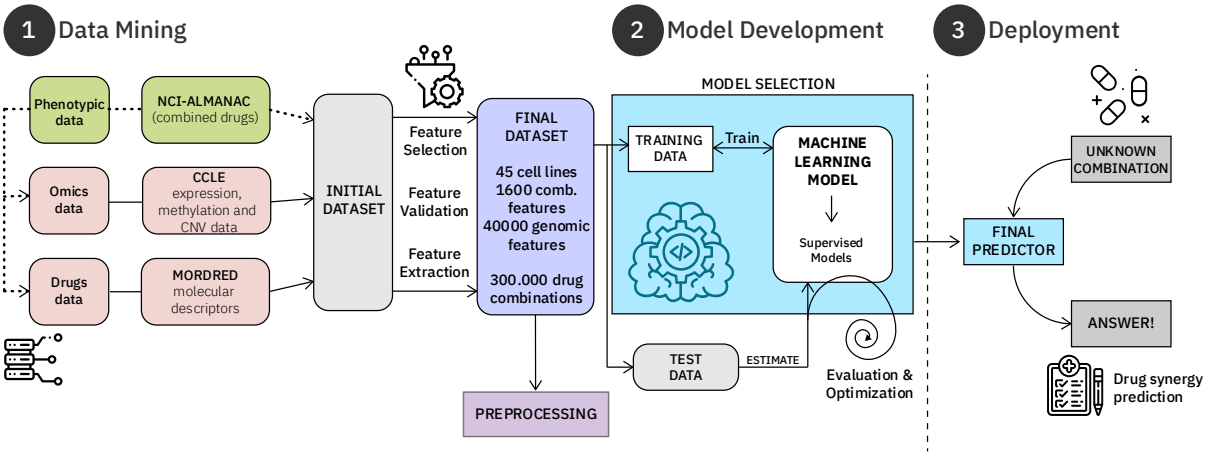



Figure 2-1: Overall workflow of the material and methods used in this work.

## 2.1.1 Data Mining

### *2.1.1.1 Raw Data Retrieval*

#### *2.1.1.1.1 Cancer Drug Combination Data*

Raw data containing the combinatory effect of drug pairs were obtained from NCI-ALMANAC via bulk-download through "https://dtp.cancer.gov/ncialmanac/" (Holbeck *et al.*, 2017). This dataset includes response data (cell killing and growth inhibition) of 104 FDA-approved drugs tested in combination across the 59 cell lines involved in the NCI-60 project (Shoemaker, 2006), making a total of 300.091 drug pair/cell line combinations. Drug sensitivity assays were performed at the NCI's Frederick National Laboratory for Cancer Research, the Stanford Research Institute, and the University of Pittsburgh. For each assay, cells were cultivated for 48 hours, in a 3 × 3 concentration matrix (three concentration values for each drug in combination). From these records, we retrieved the cell growth percentage and combination benefit scores to be used in subsequent steps. Cell growth percentage values correspond to the percentage of growth of the cell's lines in the presence of a drug combination. Combination benefit scores were calculated using a modified version of Bliss Independence (further detailed in section 3.1.1.2.1) (Bliss, 1939) (Equation 2-1).

$$CS = \sum_{p,q} Y^{A_p B_q} - Z \qquad \text{Equation 2-1}$$

where $Y^{A_p B_q}$ is the growth fraction after administration of drug $A$ at a concentration $p$ and drug $B$ at a concentration $q$; and $Z$ the control.

#### *2.1.1.1.2 Drug Physico-chemical Properties*

Each drug obtained from NCI-ALMANAC was analysed to extract their physico-chemical and structural features. The Simplified Molecular Input Line Entry System (SMILES) strings of each drug (Weininger, 1988), which were obtained by a manual query in PubChem (Kim *et al.*, 2019), were used as a mean of getting a machine-interpretable representation of those molecules. The molecules comprising atoms with unusual valence values were analysed using ChemAxon Chemicalize Tool (*Chemicalize*, 2019) and were kept according to the most common chemical species at physiological conditions. Using the Python package Mordred (Moriwaki *et al.*, 2018), we obtained for each SMILE a total of 1.825 molecular descriptors represented by numerical values, corresponding to 48 different types of features.

#### *2.1.1.1.3 Genomic Data*

Genomic features from each of the 59 cell lines, including Cell Line Annotation, Expression, CNV and Methylation, were obtained via bulk download from CCLE website (https://portals.broadinstitute.org/ccle/data) (Barretina *et al.*, 2012). OMICs data were not accessible for all cancer cell lines, so we only retrieved available information (Table 2-1). The annotation of the 48 cell lines corresponding to 9 tumour types (Blood, Brain, Breast, Colon, Kidney, Lung, Ovarian, Prostate, and Skin) originated an array of 16.394 genes. The total number of genes was then used to access the information concerning Expression, CNV and Methylation data, although, for a few genes, these data were not available (Table 2-1).

Expression data from RNA-Seq transcript quantification is normalized by RNA-Seq by Expectation Maximization (RSEM) (file: CCLE_RNAseq_rsem_genes_tpm_20180929.txt.gz) and generated an array of 14.410 gene expression measurements corresponding to 48 cell lines. Copy number variation from Affymetrix SNP6.0 Arrays was normalized by the most similar HapMap normal samples (file: CCLE_copynumber_byGene_2013-12-03.txt) originating an array of 15.367 genes from 49 cell lines. Methylation data derived by quantification of CpG islands using Reduced Representation Bisulfite Sequencing (RRBS) (file: CCLE_RRBS_tss_CpG_clusters_20181022.txt.gz) created an array of 10.021 genes from 46 cell lines. Contrarily, mutation data was programmatically retrieved via automated queries using the Selenium WebDriver Python package (Muthukadan, 2011). Fifteen kinds of mutations were retrieved (splice sites, silent, nonsense, in frame insertions, in frame deletions, frame shift insertions, frame shift deletions, *de novo* start out-of-frame, missense, start codon single nucleotide polymorphisms, nonstop, start codon insertions, stop codon insertions, stop codon deletions and start codon deletions) from a total of 16.394 genes among 48 of the 59 cell lines.

Table 2-1: Number of cell lines for type of input OMICs data.

| Type of OMICs Data | No. of Cell Lines | No. of Genes |
|---|---|---|
| **Expression** | 48 | 14.410 |
| **Copy Number Variation** | 49 | 15.367 |
| **Methylation** | 46 | 10.021 |
| **Mutations** | 48 | 16.394 |

### *2.1.1.2 Data Pre-processing*

The final dataset comprised three different types of input data: *i*) the genomic background of the 59 individual cell lines represented by the different OMICs, *ii*) the chemical properties of the drugs and iii) their corresponding combinatory effect responses. All feature columns comprising missing data were ignored in future steps. At the end of the data retrieval stage, the dataset comprised 102 drugs, 45 cell lines (a total of 226.297 samples), described using 1.244 drug-related features and 56.192 genomic descriptors (the final drugs and cell lines are displayed in Table 3 and Table 4).

Table 2-2: List of drugs comprised in the final dataset divided according to Anatomical Therapeutic Chemical (ATC) Classification (World Health Organisation, 2019) (continues in the next page).

| **Drugs Comprised in the Dataset** | | | |
|---|---|---|---|
| *Anthracyclines and related substances* | | | |
| Daunorubicin | Doxorubicin | Mitoxantrone | Valrubicin |
| *Anti-estrogens* | | | |
| Fulvestrant | Tamoxifen | | |
| *Detoxifying agents for antineoplastic treatment* | | | |
| Amifostine | Dexrazoxane | | |
| *Folic acid analogues* | | | |
| Methotrexate | Pemetrexed | Pralatrexate | |
| *Nitrogen mustard analogues* | | | |
| Bendamustine | Chlorambucil | Cyclophosphamide | Ifosfamide |
| Mechlorethamine | Melphalan | | |
| *Nitrosoureas* | | | |
| Carmustine | Lomustine | Streptozocin | Uracil mustard |
| *Other alkylating agents* | | | |
| Dacarbazine | Pipobroman | Temozolomide | Triethylenemelamine |
| *Other antineoplastic agents* | | | |
| Altretamine | Arsenic trioxide | Bortezomib | Estramustine |
| Irinotecano | Hydroxyurea | Mitotane | Pentostatin |

| | | | |
|---|---|---|---|
| Romidepsin | Topotecan | Tretinoin | Vismodegib |
| Vorinostat | | | |
| *Other cytotoxic antibiotics* | | | |
| Bleomycin | Ixabepilone | Mitomycin | Plicamycin |
| *Other immunosuppressants* | | | |
| Lenalidomide | Thalidomide | | |
| *Platinum compounds* | | | |
| Carboplatin | Cisplatin | Oxaliplatin | |
| *Podophyllotoxin derivatives* | | | |
| Etoposide | Teniposide | | |

**Drugs Comprised in the Dataset *(Cont.)***

| | | | |
|---|---|---|---|
| *Protein kinase inhibitors* | | | |
| Axitinib | Crizotinib | Dasatinib | Erlotinib |
| Everolimus | Gefitinib | Imatinib | Lapatinib |
| Nilotinib | Pazopanib | Ruxolitinib | Sorafenib |
| Sunitinib | Vandetanib | Vemurafenib | |
| *Purine analogues* | | | |
| 2-Fluoro Ara-A | Cladribine | Clofarabine | Mercaptopurine |
| Nelarabine | Thioguanine | Azacitidine | Capecitabine |
| Cytarabine | Decitabine | Floxuridine | Fluorouracil |
| Gemcitabine | | | |
| *Taxanes* | | | |
| Cabazitaxel | Docetaxel | Paclitaxel | |
| *Vinca alkaloids and analogues* | | | |
| Vinblastine | Vincristine | Vinorelbine | |
| *Other Classification* | | | |
| Abiraterone | Allopurinol | Aminolevulinic acid | Busulfan |
| Celecoxib | Dactinomycin | Imiquimod | Megestrol |
| Methoxsalen | Procarbazine | Raloxifene | Sirolimus |
| Thiotepa | Zoledronic | | |

Table 2-3: Cell lines comprised in the final dataset divided according to cancer type.

| Cancer Type | Cell Line | Cancer Type | Cell Line |
|---|---|---|---|
| Blood | HL-60(TB) | Lung | A549/ATCC |
| | K-562 | | EKVX |
| | RPMI-8226 | | HOP-62 |
| | SR | | HOP-92 |
| Brain | SF-268 | | NCI-H23 |
| | SF-295 | | NCI-H322M |
| | SF-539 | | NCI-H460 |
| | SNB-75 | | NCI-H522 |
| | U251 | Ovarian | DU-145 |
| Breast | BT-549 | | IGROV1 |
| | HS 578T | | OVCAR-3 |
| | MCF7 | | OVCAR-4 |
| | MDA-MB-231/ATCC | | OVCAR-8 |
| | MDA-MB-468 | | PC-3 |
| | T-47D | | SK-OV-3 |
| Colon | HCT-116 | Skin | LOX IMVI |
| | HCT-15 | | MALME-3M |
| | KM12 | | SK-MEL-28 |
| | SW-620 | | SK-MEL-5 |
| Kidney | 786-0 | | UACC-257 |
| | A498 | | UACC-62 |
| | ACHN | | |
| | CAKI-1 | | |
| | UO-31 | | |

#### *2.1.1.2.1 Class Assignment*

The growth percentage and combination benefit values acquired from NCI-ALMANAC were converted into binary values, representing the classification of drug combinations as synergistic or antagonistic. Each class assignment value (synergistic, antagonistic or additive effect) was obtained using five different methods (NCI-ALMANAC, Bliss, Loewe, HSA and ZIP) and evaluated independently. Classes using NCI-ALMANAC were assigned according to the mean value of the combination benefit score of each combination screening (mean of the combination score across all concentrations of the drug combination pairs) included in the raw dataset. Bliss, Loewe, HAS and ZIP scores were calculated using the R package *SynergyFinder* (He *et al.*, 2018). The binary classes, $c$ ($c \in \{0,1\}$), were assigned according to the score $s$ in each method so that if the score was positive or zero an antagonistic or additive

effect is attributed, respectively. If the score was negative, a synergistic effect is assigned (Equation 2-2).

$$c = \begin{cases} 0, & s \geq 0 \\ 1, & s < 0 \end{cases}$$ Equation 2-2

#### *2.1.1.2.2 Feature Validation and Selection*

Drug-related features were evaluated via Divisive Hierarchical Clustering using the Python package Scikit-learn (Pedregosa *et al.*, 2011). Physico-chemical features regarding the tested drugs were characterized through descriptive statistics. Features with zero variance across the full group of drugs or having missing values in any of the samples were excluded from further procedures. This proceeding reduced the number of drug-related features to 714 of the 1.244 initial descriptors distributed across 26 groups (Table 2-4).

Table 2-4: Number of features according to the type of molecular descriptor from MORDRED.

| Number of Features per Kind | | | |
|---|---|---|---|
| **Atom-bond connectivity** | 2 | **ADME** | 2 |
| **Acidity/Basicity** | 2 | **Molecular Operating Environment** | 54 |
| **Aromatic** | 2 | **Path Count** | 21 |
| **Atom Count** | 16 | **Polarizability** | 2 |
| **Autocorrelation** | 181 | **Ring Count** | 139 |
| **Bond Counts** | 9 | **Rotatable Bonds** | 1 |
| **Carbon Orbital** | 9 | **Wildman-Crippen Index** | 2 |
| **Constitutional** | 14 | **Topological Charge** | 22 |
| **Eccentric Connectivity Index** | 1 | **Topological Index** | 11 |
| **Energy State** | 158 | **Topological Polar Surface Area** | 2 |
| **Fragment Complexity** | 1 | **Walk Count** | 21 |
| **Framework** | 1 | **Weight** | 2 |
| **Hydrogen Bonds** | 2 | **Information Content** | 37 |

Genomic features were evaluated using *k*-means Agglomerative Clustering through the Python package Scikit-learn (Pedregosa *et al.*, 2011) in order to check the machines' ability to find patterns within that data. The number of clusters used in this method were chosen in line with the results of the "Elbow" Method (Appendix 1). Data was evaluated independently according to the type of genomic data (methylation, CNV, expression and mutations). Principal

Component Analysis (PCA)[9] was additionally used to independently reduce the dimensionality of genomic features.

#### *2.1.1.2.3 Feature Scaling*

Feature values were scaled across the whole dataset. For each feature type, values were normalized (Equation 2-3) and standardized (Equation 2-4).

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad \text{Equation 2-3}$$

$$x'' = \frac{x - \overline{x}}{\sigma} \quad \text{Equation 2-4}$$

$\overline{x}$ and $\sigma$ are the mean and standard variation, respectively.

## 2.1.2 Machine Learning

### *2.1.2.1 Model Architectures*

All the ML models (DNN, RF, SVM and Ensemble) were built using the Python package Scikit Learn (Pedregosa *et al.*, 2011). Each of the evaluated models was fitted with a training set corresponding to 70% of the full dataset. The remaining 30% were used as a test set for model evaluation. A grid-based search was first applied in order to find the best suited hyperparameters (learning-rate, learning-rate decay, loss functions and optimizers) by recursively testing and evaluating the created models.

### *2.1.2.2 Model Evaluation*

Different performance metrics were used to evaluate the test set of the ML models: accuracy (*acc*), precision (*prec*), recall (*rec*) (Moreira *et al.*, 2017), Area Under the Receiver Operating Characteristics Curve (ROC AUC) (Fawcett, 2002) and Mean Squared Error (MSE) (Landis and Koch, 1977). The calculation of the first four metrics implies the use of four concepts obtained by comparing the predicted outputs and the target classes: true positives (*TP*) – number of samples classified as positive in the dataset and prediction; false negatives (*FN*) – number of samples classified as positive in the dataset but classified as negative in the prediction; false positives (*FP*) - number of samples classified as negative in the dataset but classified as positive in the prediction; and true negatives (*TN*) - number of samples classified as negative in the dataset and prediction (Table 2-5) (Homenda and Pedrycz, 2018). Accuracy

[9] Principal Component Analysis (PCA) is a multivariate dimensionality reduction technique that represents several variables into new orthogonal variables called Principal Components (PC), extracting the most important information (Abdi and Williams, 2010).

measures the error rate of the model by calculating the mean of the ratio between the predicted outputs and the classes assigned in the dataset of each sample (Equation 2-5). On the other hand, precision and recall, evaluate true and false events. Precision is the fraction of detections reported by the model that were correct (Equation 2-6); recall is the fraction of true events that were detected (Equation 2-7). Both metrics can be aggregated into one single metric: F-score (Equation 2-8). The ROC AUC (Equation 2-9) is the total area under the ROC curve. On a ROC curve, recall is plotted in the *y* axis and selectivity (*sel*) (Equation 2-10) is plotted in the *x* axis. Mean Squared Error (Equation 2-11) measures the average squared difference between the estimated values and what is predicted.

Table 2-5: Confusion matrix summarizing the four types of outputs.

| | | Actual | |
|---|---|---|---|
| | | Positive | Negative |
| Predicted | Positive | True Positives | False Positives |
| | Negative | False Negatives | True Negatives |

$$acc = \frac{TP + TN}{TP + TN + FP + FN} \quad \text{Equation 2-5}$$

$$prec = \frac{TP}{TP + FP} \quad \text{Equation 2-6}$$

$$rec = \frac{TP}{TP + FN} \quad \text{Equation 2-7}$$

$$F - score = \frac{2 \cdot prec \cdot rec}{prec + rec} \quad \text{Equation 2-8}$$

$$ROC\ AUC = \int_0^1 rec\big(sel(x)\big)\, dx \quad \text{Equation 2-9}$$

$$sel = \frac{TN}{TN + FP} \quad \text{Equation 2-10}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\big(x_i - \bar{x}_i\big)^2 \quad \text{Equation 2-11}$$

## 2.2 Thorough Assessment of Membrane Protein Dimer Interfaces

In this section we addressed the approaches developed to assemble a new dataset containing information and characterization of the interfacial region of MPs called MEmbrane protein dimer Novel Structure Analyser database (MENSAdb). The overall pipeline for this section is schematized in Figure 2-2.

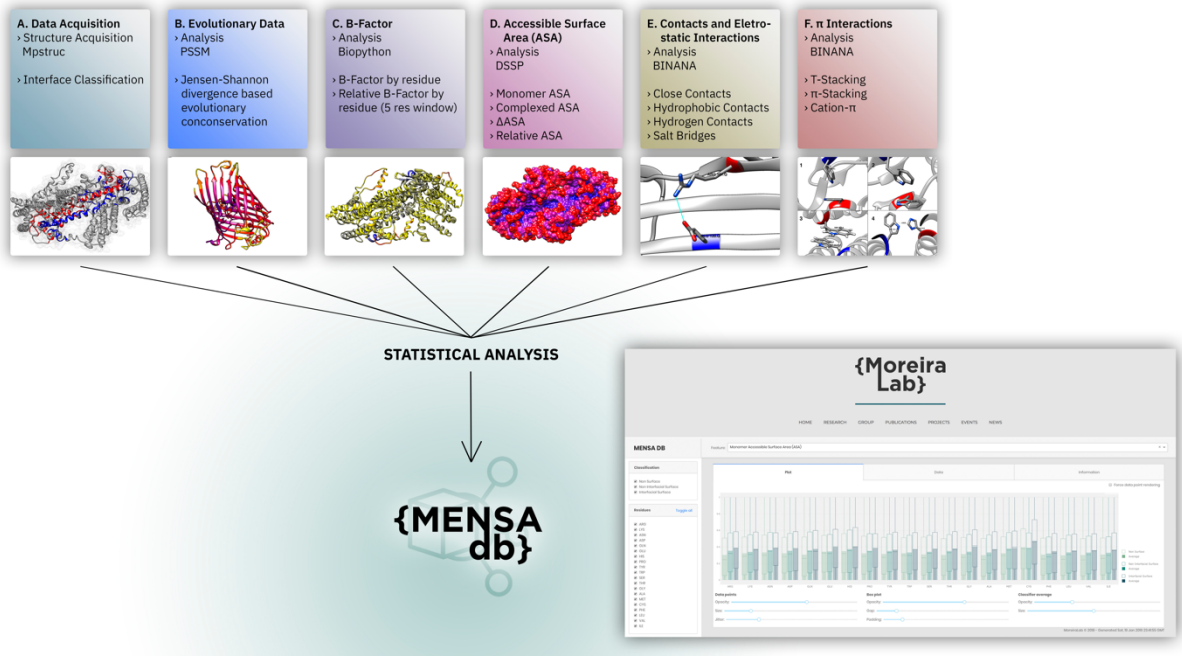



Figure 2-2: Overall representation of MENSAdb.

### 2.2.1 Raw Data Collection

Experimental structures were obtained from MPSTRUC – Membrane Proteins of Know 3D Structure (from http://blanco.biomol.uci.edu/mpstruc/) (White, 2009). This database contains 167 unique transmembrane proteins (TMs) (figures at September 2018) (including α-helix and β-barrel TMs), mostly obtained from crystal structures.

### 2.2.2 Data Pre-processing

#### *2.2.2.1 Sample Filtering*

All non-transmembrane, monomeric and monotopic proteins were discarded. Dimers in which one of the chains was a soluble protein, single MPs interacting with soluble small peptides, pores and proteins with small organic or non-organic ligands were excluded. Furthermore, structures comprising unknown residues, structures with many incomplete amino-acids, or structures with interfaces highly interacting with lipids were also disregarded. Additionally, sequences were filtered to ensure at most 35% sequence redundancy in each interface, preventing repeated complexes, using the PISCES web-server (Wang and Dunbrack, 2003). The final dataset was composed of ~63% homodimers and ~37% heterodimers.

#### 2.2.2.2 Dimer Combination Extraction

An analysis of the PDB files correspondent to the proteins contained within the dataset was performed in order to assess all dimer combinations. The final dataset was composed of 201 protein dimer combinations.

#### 2.2.2.3 Structure Standardization

PyMOL (DeLano, 2015), Modeller (Webb and Sali, 2016) and Visual Molecular Dynamics (Humphrey, Dalke and Schulten, 1996) scripts, as well as manual curation were used to identify and remove residues outside the transmembrane domain; reverse mutated non-standard amino-acids; model incomplete structures; and add hydrogens to the structures.

#### 2.2.2.4 Interface Assessment

Relative solvent accessibility (RSA) (Equation 2-12) is defined as the ratio between an amino-acid Accessible Surface Area ($ASA_{Total}$) and its corresponding area in a Gly-X-Gly peptide ($ASA_{Gly-X-Gly}$). Database of Secondary Structure assignment for all Proteins (DSSP) (Touw *et al.*, 2015) was used to calculate RSA. Residues above a 0.20 RSA cut-off were considered as surface residues (Lins, Thomas and Brasseur, 2003). From a total of 91.861 residues, 55.008 were considered surface residues after applying that method. All the residues in a pairwise distance between any atom of each chain in analysis were considered an interfacial residue, thus splitting surface residues into two classes: interfacial residues (15.277) and non-interfacial residues (39.731) (Figure 3-3).

$$RSA = \frac{ASA_{Total}}{ASA_{Gly-X-Gly}}$$ Equation 2-12

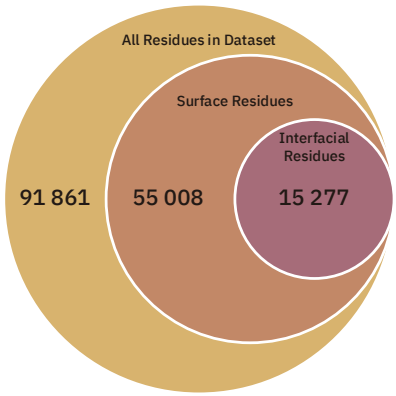



Figure 2-3: Overall distribution of the dataset.

## 2.2.3 Feature Mining

### *2.2.3.1 Evolutionary Conservation*

Evolutionary conservation of the various residues was performed using Jensen-Shannon divergence (JSD) (Lin, 1991) measure of the Position-Specific Scoring Matrix (PSSM). This PSSM matrix was calculated using PSI-BLAST (Altschul *et al.*, 1997).

### *2.2.3.2 Accessible Surface Area*

Position-Specific Scoring Matrix was used to assess ASA, RSA, measurements in the complexed $ASA_{comp}$) and monomeric $ASA_{rel}$) forms of the protein complexes. The results of those measurements were multiplied by the Sander and Rost amino-acid constants[10] (Ala: 106, Arg: 248, Asn: 157, Asp: 163, Cys:135, Gln: 198, Glu: 194, Gly: 84, His: 184, Ile: 169, Leu: 165, Lys: 205, Met: 188, Phe: 197, Pro: 136, Ser: 130, Thr: 142, Trp: 227, Tyr: 222, Val: 142) (Rost and Sander, 1994). Two additional metrics were issued to further clarify ASA values: variation of ASA from the monomeric to the complexed forms (ΔASA) (Equation 2-13) and a measurement that which allows the differentiation of residues with equal ΔASA but with different absolute monomer $ASA_{rel}$) (Equation 2-14) (Martins *et al.*, 2014).

$$\Delta ASA = ASA_{comp} - ASA_{mon} \quad \text{Equation 2-13}$$

$$ASA_{rel} = \frac{\Delta ASA}{ASA_{mon}} \quad \text{Equation 2-14}$$

### *2.2.3.3 Temperature Factor*

Biopython (Cock *et al.*, 2009), which is a python package, was used to assess the temperature factor (B-factor) of each amino-acid in analysis. An environmental B-factor measure (by issuing the mean of the B-factors of all amino-acids in a -5 – X – +5 sliding window) was also calculated to characterize the micro-environment of the proteins' residues.

### *2.2.3.4 Interfacial Interactions' Description*

BINANA-Binding Analyzer (Durrant and Mccammon, 2011), a python implemented algorithm, which characterizes protein complexes, was used to mine close[11], hydrophobic and hydrogen contacts, salt-bridges and π-interactions established between residues in the interfacial region of the protein.

[10] Amino-acid nomenclature (correspondence between full name, one letter and three letter codes) is available on page 12.

[11] Close contacts include the number of pairs of atoms formed within 2.5 and 4.0 Å distance radius.

#### *2.2.3.5 Descriptor Scaling*

Since the composition of the dataset was not equally distributed across the three classes of MPs presented here, we defined a correction factor ($C_{factor}$) (Equation 2-15) based on the concept of propensity score calculation, as presented by Huang (Huang, 2014). This factor is defined as the ratio between the frequency of occurrence of residue *i* in each one of the classes ($f_{i_{clas}}$) and the frequency of occurrence of the total number of amino-acids in that class ($f_{i_{tot}}$). The obtained MP class-specific $C_{factor}$ was used to correct the various metrics described in the Results section by multiplying them by their respective $C_{factor}$ (except for $ASA_{rel}$).

$$C_{factor} = \frac{f_{i_{class}}}{f_{i_{tot}}} \qquad \text{Equation 2-15}$$

### 2.2.4 Database Construction

#### *2.2.4.1 Descriptive Statistics*

For all plots, residues are ordered by increasing hydrophobicity based on the Kyte and Doolittle hydropathy index (Kyte and Doolittle, 1982). Descriptive statistics such as three quartiles ($Q_1$, $Q_2$ and $Q_3$), average and standard deviation were obtained using Pandas (McKinney, 2010), a Python package. All the reported p-values were calculated through SciPy (https://docs.scipy.org/), using an independent Student's $\mathcal{T}$-test. Further statistics were calculated for amino-acids sets that were split according to the hydrophilic and hydrophobic potential as: charged – Asp, Glu, Lys, Arg; positively charged – Lys, and Arg; negatively charged – Asp and Glu; polar – Ser, Thr, Asn, Gln, Tyr and His; non-polar – Ala, Val, Ile, Leu, Met, Phe and Trp; aromatic – Phe, Trp and Tyr. Cys, Gly and Pro were not included in those subsets.

#### *2.2.4.2 Web App Construction*

Data resulting from this work was made available through a web application which was built using Python's Flask-based Dash visualization framework (by Plotly (*Collaborative data science*, 2015)). The application provides graphical and tabular data formats for visual inspection and download. Real-time query features are supported by a MongoDB backend, which enables the application to query, filter and aggregate the dataset in multiple meaningful ways.

# Chapter 3. Results and Discussion

In this chapter we listed the results of the various analyses thoroughly explained in Chapter 3. The results, and corresponding critical discussion, are divided in two sections: *i*) construction of the ML model to predict the combinatory effects of anticancer drugs; *ii*) depiction of a new database describing physico-chemical features of MPs and their interfacial regions.

## 3.1 Prediction of the Combinatory Effect of Chemotherapeutic Drugs

### 3.1.1 Overview of the dataset structure

The final processed dataset comprised 1.643 features of expression values, CNV and methylation from 45 cell lines covering 9 cancer types and the corresponding combinatory effect of 102 drugs. The input data used in the training set comprised 158.408 instances (70%) while the remaining 67.889 (30%) were reserved for the test set for model evaluation.

#### *3.1.1.1 Combinatory response definition by class assessment*

Different methods (NCI-ALMANAC, HSA, Loewe, Bliss and ZIP) were used to assess the combinatory response of drugs. Concerning the obtained results, we can highlight the variability in class assessment between the different methods. This variability produced datasets with different total lengths due to the presence of missing values that forced those instances to be excluded in ML algorithms (Figure 3-1 A and Table 3-1).

Table 3-1: Overview of the constitution of the datasets according to the class assignment method.

| Method used for class assessment | No. instances of the final dataset | Synergistic Effect | Additive or Antagonist Effect |
|---|---|---|---|
| **NCI-ALAMANC** | 226.297 | 154.879 | 71.418 |
| **Loewe** | 101.064 | 61.087 | 39.977 |
| **Bliss** | 189.864 | 73.493 | 116.371 |
| **HSA** | 189.864 | 93.803 | 96.061 |
| **ZIP** | 194.972 | 101.787 | 93.185 |

Concerning the results of the class assessment, Figure 3-1 B shows that the use of HSA and ZIP models produced more balanced datasets (with the roughly same proportions of synergy- and antagonism-classified classes). However, the use of NCI-ALMANAC and Loewe models produced unbalanced datasets towards synergistic effects (2:1 and 3:2, respectively), and Bliss an unbalanced dataset towards antagonistic effects (4:7). The five methods presented in this work represent statistical approaches for calculating expected dose-response relationships in combination therapies (Fitzgerald *et al.*, 2006; Sun, Vilar and Tatonetti, 2013), and are focused exclusively on empirical observations that disregard the theoretical concepts of synergy and antagonism.

It is very time consuming to verify which model is theoretically more accurate. The models used to predict combinatory effects of drugs do not consider drug-target relationships of the drugs neither the precise target expression analysis concerning the tested cell lines. Nevertheless, although we were not able to assess the best model, the obtained results (class assessment) can successfully be used to build an ML model focused on predicting the combinatory effect of drugs, since the primary clinical objective of anti-cancer chemotherapy is a reduction in cellular proliferation leading to cell death. In agreement with this rationale and bearing in mind that ML models produce better results as data increases, the most useful method would be the one that produces more results. Consequently, the better method would be the one that assumes a more straightforward mathematical framework, since it presents a smaller chance of originating missing values, thus originating a larger dataset more suitable for ML training. For instance, the dataset obtained using Loewe additivity method, which is a more complex mathematical framework (summarized in chapter 2) that depends on fitting the effects of each drug on a dose-effect curve, originated a smaller dataset with a large quantity of missing data. On the other hand, using the method proposed by Holbeck and colleagues in NCI-ALMANAC, a simplified version of the Bliss independence method was applied, thus originating a larger dataset that does not include any missing values.

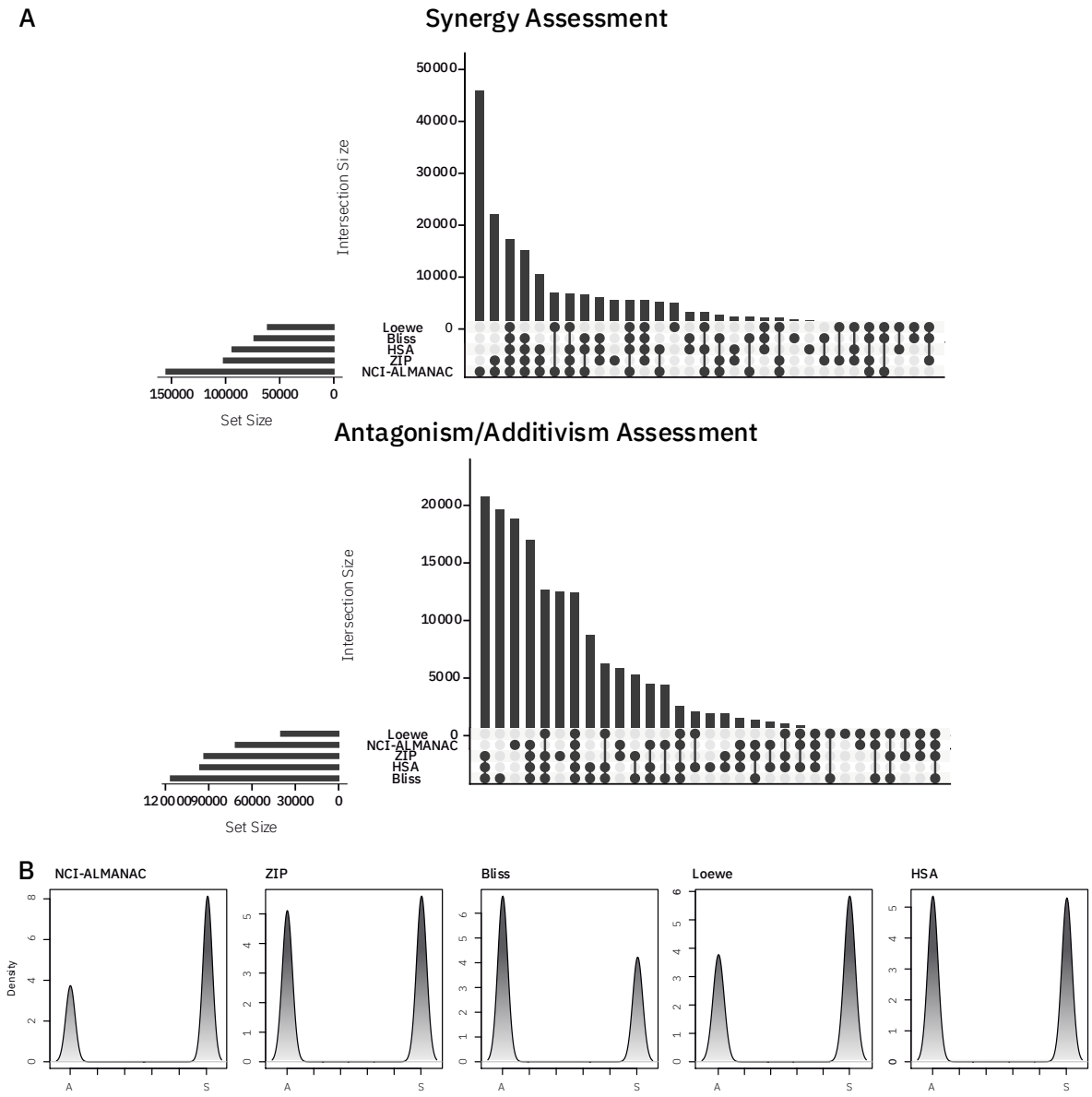



Figure 3-1: Combinatory effect classification according to the use of five different models (NCI-ALMANAC, HSA, Loewe, Bliss and ZIP). A) displays the number of classes in each dataset (top) while also showing how many of those class assignments are shared among the various datasets (bottom). B) density plots displaying the distribution of the two classes in each dataset (A represents the fraction of samples classified as Antagonistic or Additive, and S represents the fraction of samples classified as Synergistic).

### *3.1.1.2 Evaluation of drug-related features*

Evaluation of drug-related data (structural and physico-chemical features) was performed by hierarchical clustering to assess if those descriptors could describe each drug in the dataset in a meaningful way. This method exposed similarities in drug-related features across the different molecules included in the dataset, thus validating the use of these data and the computers' capability of noticing relevant patterns from them. Drugs with close structural and physico-chemical similarities appear grouped in the dendrogram (Figure 3-2): clofarabine and cladribine (adenine-like structures that have the same mechanism of action (Harned and Gaynon, 2008)); melphalan and chlorambucil (both drugs are nitrogen mustards that disrupt

DNA, preventing uncoiling and further synthesis and transcription, thus blocking DNA replication and cell proliferation (Furlanut and Franceschi, 2003; Loeber *et al.*, 2008; Rai *et al.*, 2000)); allopurinol and mercaptopurine (both drugs are purine analogues that prevent the *de novo* pathway of purine ribonucleotide synthesis (Elgemeie, 2003; Reiter *et al.*, 1983; Yang *et al.*, 2016)); ifosfamide and cyclophosphamide (both DNA alkylating agents (Zalupski and Baker, 1988)), carboplatin and oxaliplatin (both platinum-based chemotherapy drugs (Knox *et al.*, 1986)); and doxorubicin and daunorubicin (both are anthracycline antibiotics widely used in chemotherapy (Alves *et al.*, 2017; Ghirmai *et al.*, 2005)). Similarities between drugs that do not share analogous structures but are classified as belonging to the same ATC class can also be found (e.g. vincristine, vinblastine and eribulin).

The similarities identified between anticancer drugs using the physico-chemical features mined by Mordred show that this method allows the inference of key features that can be related with the therapeutic mechanism of action. In available literature, attempts to characterize drugs focus mostly on molecular fingerprints not explicitly contemplated in Mordred. For example, Morgan fingerprints (Morgan, 1965) were used by Sidorov *et al.* (Sidorov *et al.*, 2019), producing similar results to the hierarchical clustering analysis presented here. However, Morgan fingerprints present a simplistic approach when compared to Mordred descriptors as they only characterize molecules from a topological point of view (Rogers and Hahn, 2010).

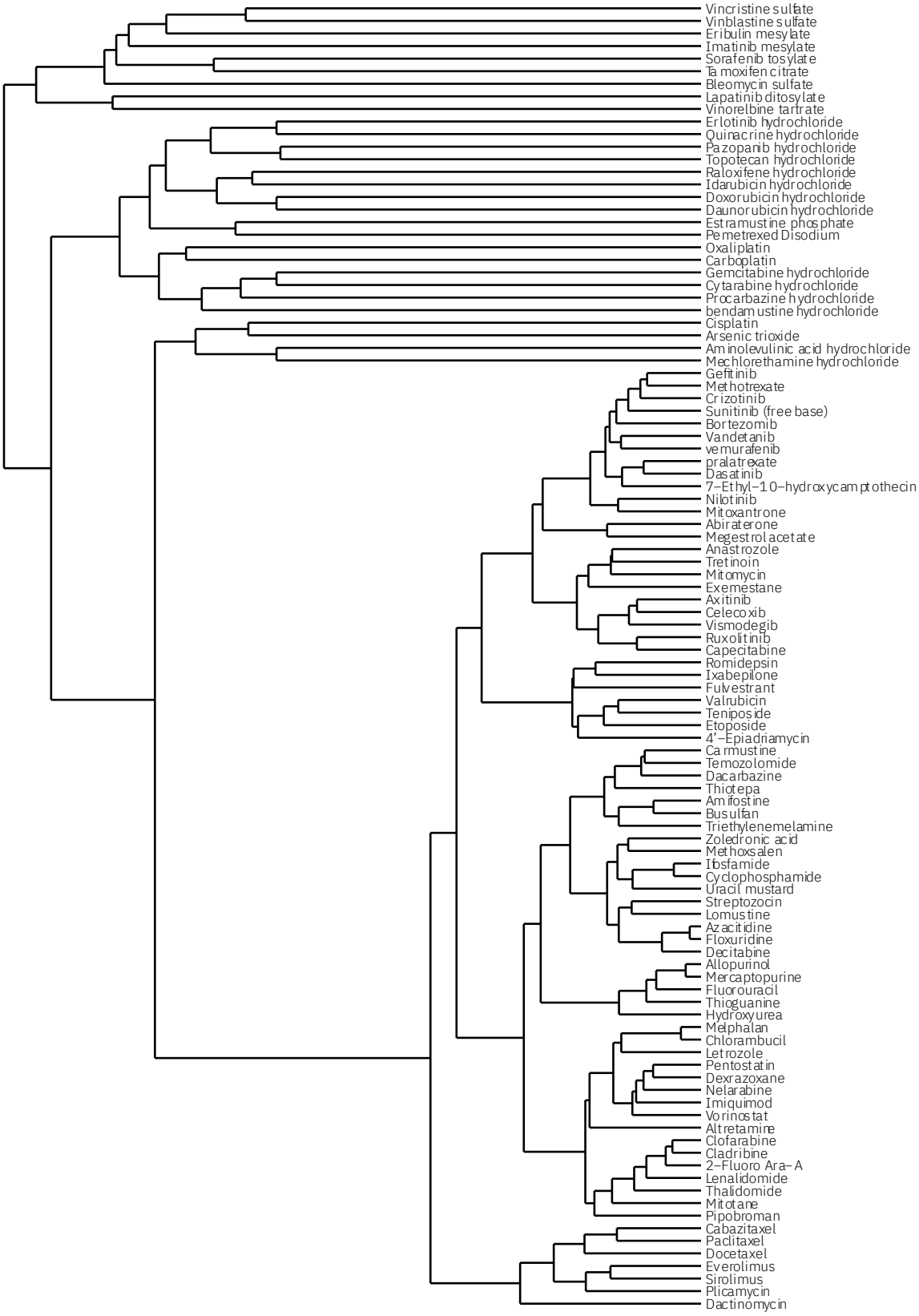



Figure 3-2: Hierarchical clustering dendrogram of all anticancer drugs included in the final dataset.

#### *3.1.1.3 Evaluation of multi-OMICs related features*

The relevance of multi-OMICs related features was assessed via independent clustering analysis of the four types of OMICs data (Expression, Methylation, CNV and Mutation) acquired before any pre-processing was made. Concerning the results of k-means clustering for evaluation of genomic data, we first assess the optimal number of clusters for each type of feature according to the elbow method. Five different clusters were chosen for analysing Mutations and Methylation; and six clusters for evaluating CNV and Expression data (Figure 3-3). In the case of mutations, 44 samples were attributed to 1 cluster; the remaining four samples were distributed across 1 cluster each. In the case of methylation, 29 samples were assigned to the first cluster, 10 to the second cluster, three samples to the third cluster, and the remaining four samples were equally distributed across the two remaining clusters. Regarding CNV, 20 samples were allocated to the first cluster, 16 samples were distributed across the second cluster, one sample to the third cluster, and the remaining 12 samples were equally distributed across the three remaining clusters. At last, expression clustering analysis allocated 20 samples in the first cluster, ten samples in the second cluster, eight samples in the third cluster, six samples in the fourth cluster, three samples in the fifth cluster and 1 sample in the remaining cluster.

Such results indicate that Expression, Methylation and CNV data allow the distinction of different cell lines from each other, as the vast majority of instances do not overlap in the analysis. On the other hand, mutation data analysed in this work were not suitable as a mean of characterizing those cells since most instances were assigned to one cluster. One of the reasons could be related with the fact that we can have changes in the way genes are switched on and off without having actual modifications in their DNA sequence, greatly supporting the significant role of epigenetics in the phenotype of cancer cells (Breindel *et al.*, 2017; Gupta *et al.*, 2019; Yuan, Norgard and Stanger, 2019). However, this conclusion cannot be taken as a staple. Other works in literature were successful using mutation data in cancer related studies. Yuan and colleagues proposed a DNN-based model for the prediction of cancer types using mutation data (Yuan *et al.*, 2016), Wood and colleagues proposed a specific SVM model for prediction of tumour-specific mutations (Wood *et al.*, 2018) and Chang and his team proposed a ML model for single-drug repositioning for cancer (Chang *et al.*, 2018). In all these works, a deep pre-processing of mutation data was performed, often by comparing cancer cell mutation data with data collected from healthy cells. This analysis often leads to the selection of driver mutations (mutations that endow selective growth advantages, and thus promote cancer development), disregarding more common passenger mutations (mutations that show no

phenotypic effects on the cancer cell) (Pon and Marra, 2015; Ushijima and Asada, 2010). The fact that, in this work, we included driver and passenger mutations coupled with the methods that were not enough to highlight major differences within that data, motivated its exclusion from our ML model development and deployment.

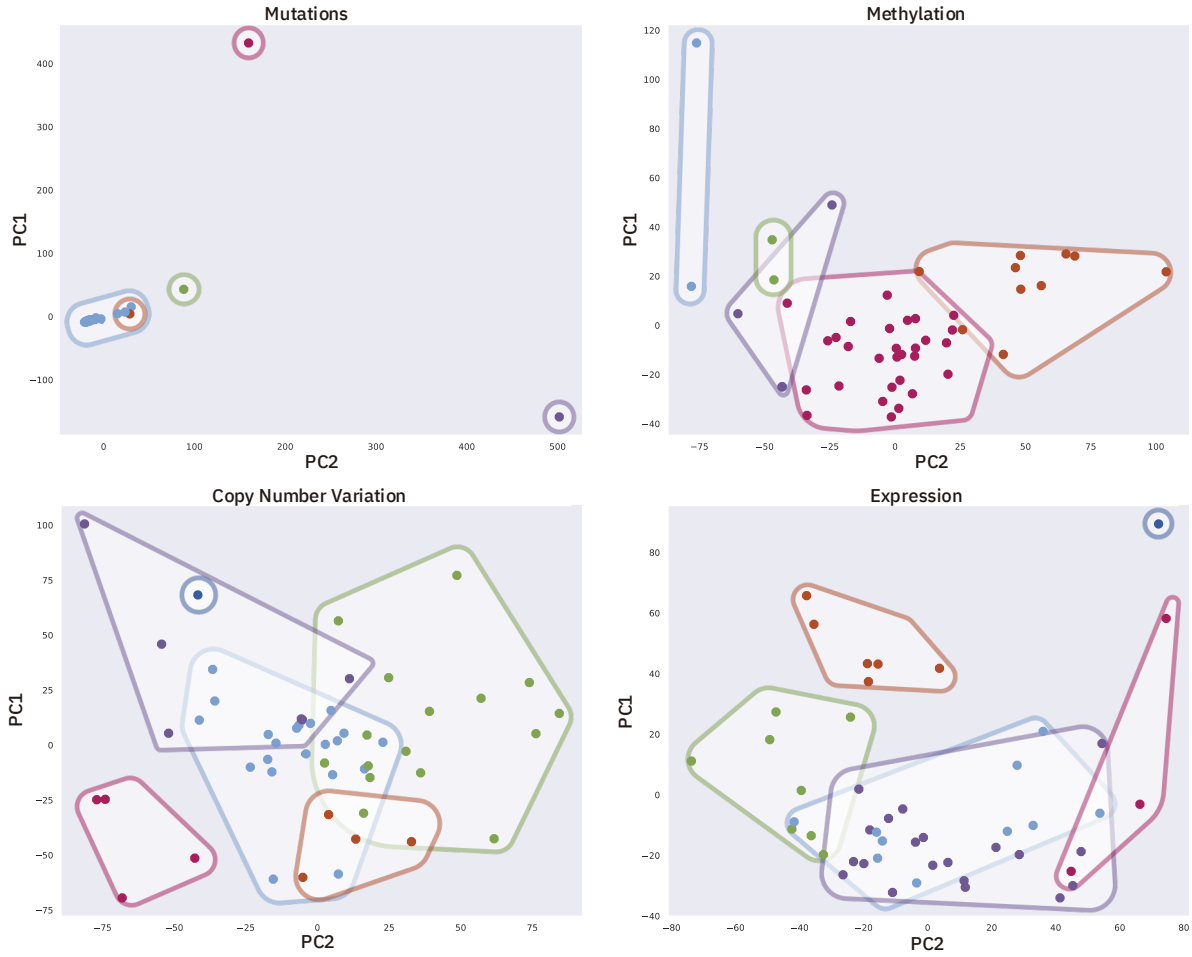



Figure 3-3: Divisive Clustering used for evaluation of cancer multi-omics related features.

### 3.1.2 Development and Evaluation of Combinatory Therapy Models

Recently, some groups are making efforts to develop models that could predict tumour drug sensitivity using classic ML methods, although mostly focused on single drug screening responses (Chang *et al.*, 2018; Chiu *et al.*, 2018; Mobadersany *et al.*, 2018). With the release of NCI-ALMANAC database (Gayvert *et al.*, 2017; Gilvary, Dry and Elemento, 2019; Sidorov *et al.*, 2019), and more recently AstraZeneca-DREAM (Bansal *et al.*, 2014; Menden *et al.*, 2019), it is now possible to integrate data from drug combination sensitivity assays. Here, we developed a model using ML methods to predict the combinatory effect of chemotherapeutic drugs based on the integration of multi-omics (expression, CNV, methylation) profiling of cell lines, drug-related features and the corresponding combinatory effects.

The development of the model was performed using three different ML approaches: Deep Neural Networks, Random Forest and SVM Classifiers. These methods were compared based on their ability to predict whether a drug combination produces synergistic, or additive/antagonistic effects (Table 3-2). Using firstly the classes obtained from the NCI-ALMANAC, our most extensive database, combination benefit score ML models were trained and evaluated. Evaluation results showed that DNN outperforms the remainder tested methods in most metrics; namely, test accuracy and MSE, with 0.73 and 0.27 respectively. On the other hand, RF beat other methods on ROC AUC and precision, with 0.67 and 0.77, respectively; SVM showed better performance in recall, and consequently in f-score, with 0.93 and 0.81, respectively. It was expected for a DNN model to outperform RFs and SVMs since it is in principle better able to handle higher volumes of data (Xue-Wen Chen and Xiaotong Lin, 2014; Zhang *et al.*, 2018), besides supporting tuning (by defining hyperparameters) that may increase its performance (Chen *et al.*, 2018). Nevertheless, none of the models shows bad evaluation results since the majority of metrics are figured above 0.70. Some recent works have also been able to develop models to predict combinatory effects of chemotherapy drugs. Preuer *et al.* presented a RF model achieving slightly better accuracy results (Preuer *et al.*, 2018). However, in this case, the use of accuracy, as an evaluation metric, can disguise the real performance of the model due to the fact of performing a mean operation across the results (Chawla, 2009). Our model presents a high value of precision, (0.77) which excludes the possibility of overfitting, outperforming results present in the literature of 0.57 at the model proposed by Preuer *et al.*

Table 3-2: Test performance metrics of different ML algorithms built using the NCI-ALMANAC dataset.

| Model | *acc* | *ROC AUC* | *MSE* | *prec* | *rec* | *f-score* |
|---|---|---|---|---|---|---|
| **DNN** | **0.7343** | 0.6676 | **0.2657** | 0.7635 | 0.8697 | 0.8131 |
| **RF** | 0.7168 | **0.6706** | 0.2832 | **0.7741** | 0.8107 | 0.7920 |
| **SVM** | 0.7154 | 0.6080 | 0.2845 | 0.7209 | **0.9336** | **0.8136** |

acc = accuracy; ROC AUC = Area Under the Receiver Operating Characteristics Curve; MSE = Mean Squared Error; prec = precision; rec = recall; Gini = Gini coefficient

Besides the original drug class assignment from NCI-ALMANAC, the results from the other four models (Bliss, HSA, Loewe and ZIP) were also independently tested using DNN (Table 3-3), as it was shown to be the best overall performing model in the results displayed above. From the five used classification models, the best performance was obtained with the

one trained against the dataset that used the combination benefit scores acquired through NCI-ALMANAC. Data from NCI-ALMANAC outperformed all the other synergy calculation methods in all metrics, empowering the No-Free-Lunch Theorem from Wolpert: "The best classifier may not be the same for all the datasets" (Meester, 2009). These performance results would seem odd results at first glance since most of the other datasets are more balanced than NCI-ALMANAC. However, the number of missing values assigned during the class calculation procedure may have caused too small datasets that do not allow the development of a good generalizing model.

Table 3-3: Test performance metrics of DNN trained with the five different synergy classification methods.

| Model | *acc* | *ROC AUC* | *MSE* | *prec* | *rec* | *f-score* |
|---|---|---|---|---|---|---|
| **Bliss** | 0.6416 | 0.5908 | 0.3584 | 0.5649 | 0.3587 | 0.4388 |
| **HSA** | 0.6451 | 0.6255 | 0.3549 | 0.6111 | 0.4843 | 0.5404 |
| **Loewe** | 0.6492 | 0.5981 | 0.3508 | 0.6636 | 0.8474 | 0.7443 |
| **NCI-ALMANAC** | **0.7343** | **0.6676** | **0.2657** | **0.7635** | **0.8697** | **0.8131** |
| **ZIP** | 0.6232 | 0.6214 | 0.3768 | 0.6485 | 0.5098 | 0.5709 |

acc = accuracy; ROC AUC = Area Under the Receiver Operating Characteristics Curve; MSE = Mean Squared Error; prec = precision; rec = recall; Gini = Gini coefficient

The evaluation results obtained from the three ML models (DNN, RF and SVM) (Table 3-2) suggest that combining the three methods to build one standalone model would create a better performing model. A hard-voting ensemble model (with a similar structure to the ensemble decision trees that compose RF) aggregating the three ML models was trained and tested with the NCI-ALMANAC dataset. This new model obtained slightly better results, achieving 0.7404, 0.9051 and 0.8226 in accuracy, precision and f-score, respectively (Table 3-4). In fact, previous studies (Singh, Rana and Singh, 2018) using ensemble models, report better results when applying those methods in comparison to the individual algorithms used for model development (Moreira *et al.*, 2017).

Table 3-4: Evaluation of the ensemble model combining DNN, RF and SVM models in Table 6.

| Model | *acc* | *ROC AUC* | *MSE* | *prec* | *rec* | *f-score* |
|---|---|---|---|---|---|---|
| **Ensemble** | **0.7404** | 0.6593 | 0.2596 | 0.7539 | **0.9051** | **0.8226** |

acc = accuracy; ROC AUC = Area Under the Receiver Operating Characteristics Curve; MSE = Mean Squared Error; prec = precision; rec = recall; Gini = Gini coefficient

An in-depth assessment of the DNN model accuracy (trained using the NCI-ALMANAC dataset) per ATC classification and per type of cell line was performed. Drugs classified as "Other Alkylating Agents" had a best overall test performance (0.80 ± 0.08), followed by "Nitrogen mustard analogues" (0.76 ± 0.04), "Nitrosoureas" (0.75 ± 0.04) and "Purine analogues" (0.75 ± 0.03). On the other hand, "Anthracyclines and related substances" (0.66 ± 0.06), "Other cytotoxic antibiotics" (0.66 ± 0.06) and "Platinum compounds" (0.64 ± 0.08) returned the lowest accuracy results. Evaluation results per ATC classification could be in-depth analysed in Supplementary Table 2. Concerning the results of model accuracy per cell, colon cancer cells presented the best overall accuracy (0.74 ± 0.03), followed by kidney cancer cells (0.73 ± 0.02) and ovarian cancer cells (0.73 ± 0.02). In opposition, blood cancer cells (in this case, representative of haematological malignancies within the dataset) present the worst accuracy evaluation (0.70 ± 0.04) (Figure 3-4). This result denotes the significant differences between solid tumour cells and their haematological counterparts, assuming a possible source of entropy within the dataset that may prevent the model from achieving more significant results. The model may also interpret the smaller number of samples of haematological malignancies cell lines and their genomic background as noise, which may be also contributing for the obtained results. In fact, in contrast with solid tumours, that are abnormal mass of cells, haematological malignancies that move through blood and lymph, present specific genomic backgrounds and interactions with the tumour microenvironment (Zhou, 2005). A possible solution would be to handle each type of cancer cells independently, splitting the data into two different datasets. Moreover, similar results can also be found in the literature. Differences between models' performance according to the type of cancer from which the data are acquired (from solid or haematological malignancies) were also observed by the model proposed by Sidorov *et al.* (Sidorov *et al.*, 2019). Further cell line evaluation results per type of cancer cell line can be in-depth analysed in Supplementary Table 3.

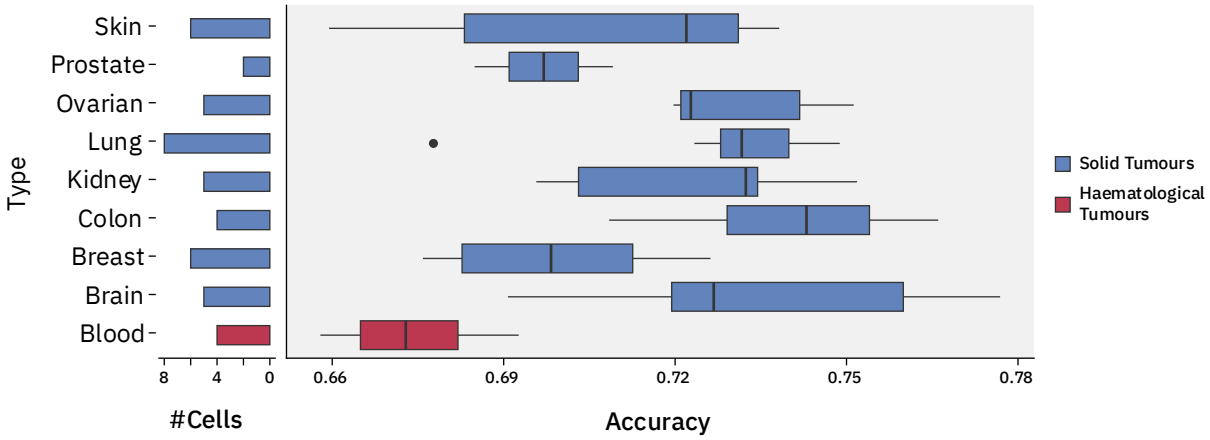



Figure 3-4: Accuracy per type of cancer cell line.

## 3.2 Assessment of Membrane Protein Dimer Interface Characteristics

The objective of this part of the research work was to create a comprehensive real-time web-application exposing a broad array of fundamental structural and physico-chemical features of a curated collection of membrane protein dimer structures and their interfacial regions, the MEmbrane protein dimer Novel Structure Analyser database (MENSAdb). Due to the importance of MPs as targets of more than half of all current drugs on the market, the mining of these new data describing MPs interfaces could provide additional features that can be added to the combinatory effect model presented and discussed in section 4.1 to further enhance the performance.

### 3.2.1 Membrane protein dimer composition

The overall residue distribution in Figure 3-5 shows that MPs dimers have a higher content of hydrophobic and aromatic residues, such as leucine, alanine, valine, glycine, isoleucine and phenylalanine that account for 55% of all detected residues. This high hydrophobic content was also previously reported in several studies (Eilers *et al.*, 2002; Saidijam, Azizpour and Patching, 2018; Ulmschneider and Sansom, 2001).

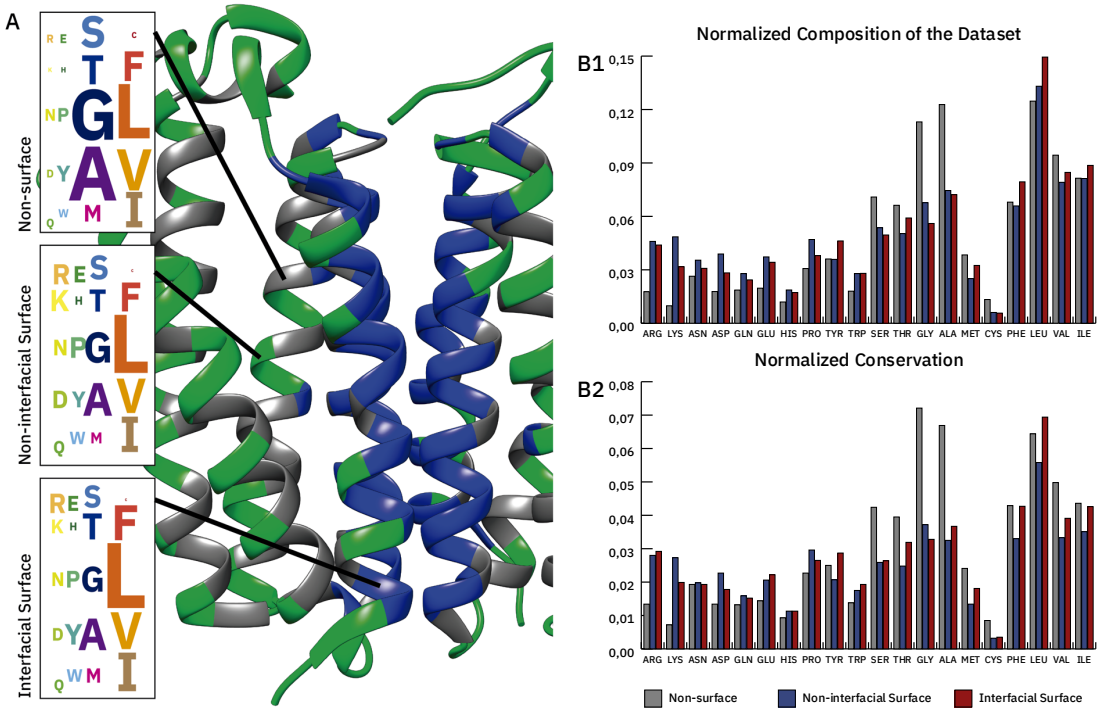



Figure 3-5: Structural and physico-chemical properties of MPs and their interactions. (A) – residue distribution of the translocator membrane protein (PDBid: 4UC1) from *Rhodobacter sphaeroides* (Li *et al.*, 2015). (B1) – residue composition of the dataset. (B2) – evolutionary conservation scores. Amino-acid nomenclature (correspondence between full name, one letter and three letter codes) is available on page 12.

The overall distribution of individual residues of MPs by amino acid type (Figure 3-5 B1 and B2) shows that GAS residues (Glycine, Alanine, Serine) (Zhang *et al.*, 2015) are particularly enriched at the MPs non-surface. These small residues are the strong driving force for membrane folding (Zhang *et al.*, 2009). As expected, charged residues are excluded from the MPs non-surface. The propensities for charged and polar residues at interfaces are intermediate between those for non-surface and non-interface surfaces. Residue distributions are in close agreement with several studies demonstrating that Protein-Protein Interactions (PPIs) are mostly hydrophobic (e.g., leucine, isoleucine) in nature, with some aromatic residues (e.g., phenylalanine and tyrosine) and yield a buried non-polar surface area (Ulmschneider and Sansom, 2001; Yan *et al.*, 2008).

Evolutionary conservation of protein sequences is a key feature for understanding what are the functionally and structurally important residues in protein-protein interfaces. We used JSD dissimilarity score, in which values close to 0 mean a similar distribution whereas scores of 1 corresponds to totally discordant distributions. Figure 3-5 B2 reveals that the highest JSD corrected values differences are for the more conserved GAS residues in the non-surface, and the non-polar residues in the interface.

Average B-factor[12] values (by residue and using a five-residue window) measure the fluctuation of an atom around its mean position. Various authors have suggested that for soluble PPIs lower B-factors values for interfacial residues are indicative of lower flexibility (Chakravarty *et al.*, 2015; Jones and Thornton, 1995; Liu, Jiang and Zhou, 2010). We observed a decrease in corrected B-factor values of the interfacial residues compared to the non-interfacial surface ones (5.71 $\pm$ 6.10 vs 6.25 $\pm$ 6.16; p-value=$5.12\times10^{-20}$), putting their average closer to the non-surface MP residues (6.02 $\pm$ 5.69). Also, interfacial surface and non-surface positively charged residues are the most dissimilar (3.74 $\pm$ 2.86 vs 1.19 $\pm$ 0.96; p-value=$2.63\times10^{-140}$). The same holds true for environmental B-factor. These observations agree with findings attained for soluble PPIs (Chakravarty *et al.*, 2015).

### 3.2.2 Characteristics of interfacial residues

Identification and characterization of critical features of membrane dimer PPIs can provide important clues to pinpoint residues or interactions, important for drug development and repurposing. For this, additional interfacial structural characteristics were quantified to better understand MPs dimers. Concerning the intermolecular atomic contacts per amino-acid type, we observed that the aromatic residues (corrected contacts at 4 Å: 0.56 ± 0.61) are much more prone to establish close contacts at short distance than other residues. Arg was also highlighted in our results (corrected contacts at 4 Å: 0.75 ± 0.82).

Additionally, although MP residues reside in an apolar (low dielectric) environment (Lomize *et al.*, 2007; Zhang, Witham and Alexov, 2011), both salt-bridges between charged residues and hydrogen-bonds through almost all amino-acids are common to stabilize the interface and promote complex formation. Hydrogen-bonds measured here involving both side-chains and backbone are particularly important for polar (corrected according to Equation 2-15: 0.01±0.03) and charged residues (corrected values: 0.01±0.03) but also for aromatic ones (corrected values: 0.01±0.02), in particular Tyr (corrected values: 0.01±0.03) and Trp (corrected values: 0.01±0.01).

### 3.2.3 Web App Accessibility

Data resulting from this work are available through the MENSAdb Web App, accessible at http://www.moreiralab.com/resources/mensadb, where further information can be accessed.

[12] This metric is exclusive to crystal-based structures.

# Chapter 4. Conclusions and Future Perspectives

The first objective of this research project was to apply AI techniques, particularly ML algorithms, to predict the combinatory effect of anticancer drugs using structural and physico-chemical information from each molecule and biological knowledge acquired through high-throughput OMICs technologies from cancer cell lines. For the first part of this work, NCI-ALMANAC provided raw drug combination data, and CCLE presented CNV, Expression, Methylation and Mutation data from cancer cell lines. Concerning the characterization of the drugs contained in NCI-ALMANAC, this information was mined in-house using Mordred. After pre-processing, clustering methods were used to evaluate the quality and relevance of that data, highlighting the important role of genomic data for the characterization of cell phenotypes.

Four ML models were built and tested: a DNN, a RF, a SVM, and an Ensemble model combining the last three. The best performing model, the Ensemble model, achieved 0.74 accuracy, 0.75 precision and 0.90 recall, attaining results capable of predicting new combinations for chemotherapy by performing drug screening assays and eliminating less advantageous candidates, especially in the case of solid tumours. It was also possible to understand how numeric vectors describing such intangible entities like small molecules or genomic information of living cells can allow so abstract things as computers to draw human-like conclusions from biological data. According to the obtained results, we can conclude that our model has a lot of potential for implementation in the development of future drug combinations, outperforming others in literature in terms of reliability and generalizability.

Better results may, nevertheless, be achieved. Firstly, by splitting the dataset into solid tumours and haematological malignancies, and training independently the model we can achieve better results. Secondly, by performing further experiments with a larger dataset that

comprises more drugs and more cell lines, will allow to better tune the proposed ML models. Thirdly, maintaining the same dataset, other directions can also be explored in future research, namely by using other kinds of data, like other types of OMICs (e.g., proteomics and metabolomics), action mechanism pathways and target-related data. We began to track that path towards producing new target-related features by creating MENSAdb, a database, and a user accessible Web App, containing data describing the interfacial region of MPs - the molecular target of most FDA-approved drugs. Furthermore, new ML architectures, like graph-based data and neural networks models (Bui, Ravi and Ramavajjala, 2017; Duvenaud *et al.*, 2015; Ramazzotti *et al.*, 2019), can also be used to pass knowledge into the computer in new and more comprehensive ways. Further testing to the existing model can also be achieved by performing a screening of currently unknown combinatory therapeutic strategies and an *in vitro* and *in vivo* assessment of the best candidates from that procedure, acting as true biological validation of the model herein presented.

This work originated one publicly available pre-print article (available at https://arxiv.org/abs/1902.02321) (Matos-Filipe *et al.*, 2019) that will be submitted in the near future, a peer-reviewed book chapter (available at https://link.springer.com/protocol/10.1007/978-1-4939-9161-7_21) (Preto *et al.*, 2019) and a conference proceeding presented at the *2018 Meeting of Young Structural Computational Biology Researchers*. I am also involved in two peer-reviewed book-chapters, "Prediction and targeting of GPCR oligomer interfaces" and "Deep Learning: the nodes of biological data on the edge of technology", to be published in *Progress in Molecular Biology and Translational Science* (Elsevier) and *Methods in Molecular Biology* (Springer Science), respectively, in the end of 2019. Additionally, I am also involved in a peer-reviewed original research article as co-first author: "Membrane protein dimer Novel Structure Analyser (MENSA)", to be published by the end of this year.

# References


AALST, W. M. P. VAN DER *et al.* - Process mining: a two-step approach to balance between underfitting and overfitting. **Software & Systems Modeling**. ISSN 1619-1366. 9:1 (2010) 87–111. doi: 10.1007/s10270-008-0106-z.

ABRANTES-METZ, Rosa M.; ADAMS, Christopher; METZ, Albert D. - Pharmaceutical Development Phases: A Duration Analysis. **SSRN Electronic Journal**. ISSN 1556-5068. (2004). doi: 10.2139/ssrn.607941.

AKBANI, Rehan *et al.* - A pan-cancer proteomic perspective on The Cancer Genome Atlas. **Nature Communications**. ISSN 2041-1723. 5:1 (2014) 3887. doi: 10.1038/ncomms4887.

ALTSCHUL, Stephen F. *et al.* - Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. **Nucleic Acids Research**. ISSN 03051048. (1997). doi: 10.1093/nar/25.17.3389.

ALVES, Ana Catarina *et al.* - Daunorubicin and doxorubicin molecular interplay with 2D membrane models. **Colloids and Surfaces B: Biointerfaces**. ISSN 09277765. 160 (2017) 610–618. doi: 10.1016/j.colsurfb.2017.09.058.

ASHBURN, Ted T.; THOR, Karl B. - Drug repositioning: Identifying and developing new uses for existing drugs. **Nature Reviews Drug Discovery**. ISSN 14741776. 3:8 (2004) 673–683. doi: 10.1038/nrd1468.

AUFFRAY, Charles *et al.* - Making sense of big data in health research: Towards an EU action plan. **Genome Medicine**. ISSN 1756-994X. 8:1 (2016) 71. doi: 10.1186/s13073-016-0323-y.

BALMAIN, Allan; GRAY, Joe; PONDER, Bruce - The genetics and genomics of cancer. **Nature Genetics**. ISSN 1061-4036. 33:S3 (2003) 238–244. doi: 10.1038/ng1107.

BANSAL, Mukesh *et al.* - A community computational challenge to predict the activity of pairs of compounds. **Nature Biotechnology**. ISSN 1087-0156. 32:12 (2014) 1213–1222. doi: 10.1038/nbt.3052.

BARRETINA, Jordi *et al.* - The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. **Nature**. ISSN 0028-0836. 483:7391 (2012) 603–607. doi: 10.1038/nature11003.

BECKER, Suzanna; PLUMBLEY, Mark - Unsupervised neural network learning procedures for feature extraction and classification. **Applied Intelligence**. ISSN 0924-669X. 6:3 (1996) 185–203. doi: 10.1007/BF00126625.

BEHZAD, Mohsen *et al.* - Generalization performance of support vector machines and neural networks in runoff modeling. **Expert Systems with Applications**. ISSN 0957-4174. 36:4 (2009)

7624–7629. doi: 10.1016/J.ESWA.2008.09.053.

BENNETT, Simon T. *et al.* - Toward the $1000 human genome. **Pharmacogenomics**. ISSN 1462-2416. 6:4 (2005) 373–382. doi: 10.1517/14622416.6.4.373.

BERMAN, Helen; HENRICK, Kim; NAKAMURA, Haruki - Announcing the worldwide Protein Data Bank. **Nature Structural Biology**. ISSN 10728368. (2003). doi: 10.1038/nsb1203-980.

BISHOP, Christopher M. - **Pattern Recognition and Machine Learning**. 2. ed. New York : Springer Science+Business Media, LLC, 2006. ISBN 0387310738.

BLISS, Chester Ittner - The toxicity of poisons applied jointly. **Annals of Applied Biology**. ISSN 17447348. 26 (1939) 585–615. doi: 10.1111/j.1744-7348.1939.tb06990.x.

BOUNTRA, Chas; LEE, Wen Hwa; LEZAUN, Javier - **A New Pharmaceutical Commons: Transforming Drug Discovery**. (2017)

BOZIC, Ivana *et al.* - Evolutionary dynamics of cancer in response to targeted combination therapy. **eLife**. ISSN 2050-084X. 2 (2013). doi: 10.7554/elife.00747.

BRAMER, Max - Avoiding Overfitting of Decision Trees. In **Principles of Data Mining**. London : Springer London, 2007. ISBN 9781447148845. p. 119–134.

BRAVO-MERODIO, Laura *et al.* - -Omics biomarker identification pipeline for translational medicine. **Journal of Translational Medicine**. ISSN 1479-5876. 17:1 (2019) 155. doi: 10.1186/s12967-019-1912-5.

BRAY, F. *et al.* - **Cancer Incidence in Five Continents, Vol. XI (electronic version)** [On Line], atual. 2017. [Accessed 28 mar. 2019]. Available at WWW:<URL:http://ci5.iarc.fr>.

BRAY, Freddie *et al.* - Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. **CA: A Cancer Journal for Clinicians**. ISSN 00079235. 68:6 (2018) 394–424. doi: 10.3322/caac.21492.

BREINDEL, Jerrica L. *et al.* - Epigenetic Reprogramming of Lineage-Committed Human Mammary Epithelial Cells Requires DNMT3A and Loss of DOT1L. **Stem Cell Reports**. ISSN 22136711. 9:3 (2017) 943–955. doi: 10.1016/j.stemcr.2017.06.019.

BRODY, Tom - Drug Class Analysis. In **FDA's Drug Review Process and the Package Label**. Academic Press, 2018. ISBN 978-0-12-814647-7. p. 441–511.

BUI, Thang D.; RAVI, Sujith; RAMAVAJJALA, Vivek - Neural Graph Machines: Learning Neural Networks Using Graphs. **arXiv**. (2017) 9. doi: 1703.04818.

**Cancer prevention and control in the context of an integrated approach** - [On Line], [Accessed 30 mar. 2019] Geneva : World Health Assembly, 2017. Available at WWW:<URL:http://www.who.int/nmh/events/2015/technical-note-en.pdf?ua=1>

**Cancer Research UK** - [On Line], atual. 2019. [Accessed 14 may. 2019]. Available at WWW:<URL:https://www.cancerresearchuk.org/>.

CHABNER, Bruce A.; ROBERTS, Thomas G. - Chemotherapy and the war on cancer. **Nature Reviews Cancer**. ISSN 1474-175X. 5:1 (2005) 65–72. doi: 10.1038/nrc1529.

CHAKRAVARTY, Devlina *et al.* - Changes in protein structure at the interface accompanying complex formation. **IUCrJ**. ISSN 20522525. (2015). doi: 10.1107/S2052252515015250.

CHANEY, Stephen G. *et al.* - Protein interactions with platinum-DNA adducts: From structure to function. **Journal of Inorganic Biochemistry**. ISSN 0162-0134 (2004). doi: https://doi.org/10.1016/j.jinorgbio.2004.04.024.

CHANG, Yoosup *et al.* - Cancer Drug Response Profile scan (CDRscan): A Deep Learning Model That Predicts Drug Effectiveness from Cancer Genomic Signature. **Scientific Reports**. ISSN 20452322. 8:1 (2018) 1–11. doi: 10.1038/s41598-018-27214-6.

CHAWLA, Nitesh V. - Data Mining for Imbalanced Datasets: An Overview. In **Data Mining and Knowledge Discovery Handbook**. Boston, MA : Springer US, 2009. ISBN 9780387098234. p. 875–886.

CHEMAXON. - **Chemicalize** [On Line] (2019). Available at WWW:<URL: https://chemicalize.com/>

CHEN, Hongming *et al.* - The rise of deep learning in drug discovery. **Drug Discovery Today**. ISSN 1359-6446. 23:6 (2018) 1241–1250. doi: 10.1016/J.DRUDIS.2018.01.039.

CHEN, Lan; MALHOTRA, Anshoo - Combination Approach: the Future of the War Against Cancer. **Cell Biochemistry and Biophysics**. ISSN 1085-9195. 72:3 (2015) 637–641. doi: 10.1007/s12013-015-0549-0.

CHEN, Yue *et al.* - Synthesis and evaluation of a technetium-99m-labeled diethylenetriaminepentaacetate-deoxyglucose complex ([ 99m Tc]-DTPA-DG) as a potential imaging modality for tumors. **Applied Radiation and Isotopes**. 64 (2006) 342–347. doi: 10.1016/j.apradiso.2005.08.004.

CHIU, Yu-Chiao *et al.* - Predicting drug response of tumors from integrated genomic profiles by deep neural networks. **arXiv**. ISSN 8750-7587. (2018). doi: 10.1152/japplphysiol.01617.2011.

CHOU, Ting-Chao - Drug Combination Studies and Their Synergy Quantification Using the Chou-Talalay Method. **Cancer Research**. ISSN 0008-5472. 70:2 (2010) 440–446. doi: 10.1158/0008-5472.CAN-09-1947.

CHOU, Ting-Chao; TALALAY, Paul - Quantitative analysis of dose-effect relationships: the combined effects of multiple drugs or enzyme inhibitors. **Advances in Enzyme Regulation**. ISSN 0065-2571. 22 (1984) 27–55. doi: 10.1016/0065-2571(84)90007-4.

COCK, P. J. A. *et al.* - Biopython: freely available Python tools for computational molecular biology and bioinformatics. **Bioinformatics**. ISSN 1367-4803. 25:11 (2009) 1422–1423. doi: 10.1093/bioinformatics/btp163.

PLOTLY - Collaborative data science [On Line] Montreal, QC. (2015). Available at WWW:<URL: https://plot.ly/>

**Comprehensive Cancer Information - National Cancer Institute** - [On Line], atual. 2019. [Accessed 14 may. 2019]. Available at WWW:<URL:https://www.cancer.gov/>.

CORTES, Corinna; VAPNIK, Vladimir - Support-Vector Networks. **Machine Learning**. ISSN 15730565. 20:3 (1995) 273–297. doi: 10.1023/A:1022627411411.

DAVIS, Mark E.; CHEN, Zhuo; SHIN, Dong M. - Nanoparticle therapeutics: an emerging treatment modality for cancer. **Nature Reviews Drug Discovery**. ISSN 1474-1776. 7:9 (2008) 771–782. doi: 10.1038/nrd2614.

DEBNATH, Mousumi *et al.* - Omics Technology. In **Molecular Diagnostics: Promises and Possibilities**. Dordrecht : Springer Netherlands, 2010. ISBN 9789048132614. p. 11–31.

SCHRÖDINGER, LLC - **The PyMOL Molecular Graphics System** [On Line]. New York, NY, (2019) Available at WWW:<URL: https://www.schrodinger.com/>

DIETTERICH, Thomas G. - Ensemble Methods in Machine Learning. Springer, Berlin, Heidelberg, 2000. ISBN 9783540450146. p. 1–15.

DIMASI, Joseph A. - Risks in new drug development: Approval success rates for investigational drugs. **Clin Pharmacol Ther**. 69 (2001) 297–307. doi: 10.1067/mcp.2001.115446.

DIMASI, Joseph A.; GRABOWSKI, Henry G.; HANSEN, Ronald W. - Innovation in the pharmaceutical industry: New estimates of R&D costs. **Journal of Health Economics**. ISSN 18791646. 47 (2016) 20–33. doi: 10.1016/j.jhealeco.2016.01.012.

DOGTEROM, Marileen; KOENDERINK, Gijsje H. - Actin–microtubule crosstalk in cell biology. **Nature Reviews Molecular Cell Biology**. ISSN 1471-0072. 20:1 (2019) 38–54. doi: 10.1038/s41580-018-0067-1.

DUGGER, Sarah A.; PLATT, Adam; GOLDSTEIN, David B. - Drug development in the era of precision medicine. **Nature Reviews Drug Discovery**. ISSN 1474-1776. 17:3 (2017) 183–196. doi: 10.1038/nrd.2017.226.

DURRANT, Jacob D.; MCCAMMON, J. Andrew - BINANA: A Novel Algorithm for Ligand-Binding Characterization. (2011). doi: 10.1016/j.jmgm.2011.01.004.

DUVENAUD, David *et al.* - Convolutional Networks on Graphs for Learning Molecular Fingerprints. **arXiv**. (2015). doi: http://arxiv.org/abs/1509.09292.

EILERS, Markus *et al.* - Comparison of helix interactions in membrane and soluble α-bundle proteins. **Biophysical Journal**. ISSN 00063495. (2002). doi: 10.1016/S0006-3495(02)75613-0.

ELBEHERY, Ali H. A.; AZZAZY, Hassan M. E. - Nanoparticle-based detection of cancer-associated RNA. **Wiley Interdisciplinary Reviews: Nanomedicine and Nanobiotechnology**. ISSN 19395116. 6:4 (2014) 384–397. doi: 10.1002/wnan.1266.

ELGEMEIE, Galal - Thioguanine, Mercaptopurine: Their Analogs and Nucleosides as Antimetabolites. **Current Pharmaceutical Design**. ISSN 13816128. 9:31 (2003) 2627–2642. doi: 10.2174/1381612033453677.

ESTIVILL-CASTRO, Vladimir - Why so many clustering algorithms. **ACM SIGKDD Explorations Newsletter**. ISSN 19310145. 4:1 (2002) 65–75. doi: 10.1145/568574.568575.

FAN, Wenpei *et al.* - Nanotechnology for Multimodal Synergistic Cancer Therapy. **Chemical reviews**. (2017). doi: 10.1021/acs.chemrev.7b00258.

FAWCETT, T. - Using rule sets to maximize ROC performance. In **Proceedings 2001 IEEE International Conference on Data Mining**. IEEE Comput. Soc, 2002. ISBN 0-7695-1119-8

**FDA Approved Drugs in Oncology** - [On Line], atual. 2019. [Accessed 1 apr. 2019]. Available at WWW:<URL:https://www.centerwatch.com/drug-information/fda-approved-drugs/therapeutic-area/12/oncology>.

FENG, Chunlai *et al.* - Gene Expression Data Based Deep Learning Model for Accurate Prediction of Drug-induced Liver Injury in Advance. **Journal of Chemical Information and Modeling**. ISSN 1549-9596. (2019) acs.jcim.9b00143. doi: 10.1021/acs.jcim.9b00143.

FERLAY, J. *et al.* - **Global Cancer Observatory: Cancer Today** [On Line], atual. 2018. [Accessed 19 mar. 2019]. Available at WWW:<URL:https://gco.iarc.fr/today/>.

FERLIGOJ, Anuška; BATAGELJ, Vladimir - Some types of clustering with relational constraints. **Psychometrika**. ISSN 0033-3123. 48:4 (1983) 541–552. doi: 10.1007/BF02293878.

FITZGERALD, Jonathan B. *et al.* - Systems biology and combination therapy in the quest for clinical efficacy. **Nature Chemical Biology**. ISSN 1552-4450. 2:9 (2006) 458–466. doi: 10.1038/nchembio817.

FOO, Jasmine; MICHOR, Franziska - Evolution of acquired resistance to anti-cancer therapy. **Journal of theoretical biology**. ISSN 1095-8541. 355 (2014) 10–20. doi: 10.1016/j.jtbi.2014.02.025.

FOUCQUIER, Julie; GUEDJ, Mickael - Analysis of drug combinations: current methodological landscape. **Pharmacology research & perspectives**. ISSN 2052-1707. 3:3 (2015) e00149. doi: 10.1002/prp2.149.

FRALEY, Chris; RAFTERY, AE E. - How many clusters? Which clustering method? Answers via model-based cluster analysis. **The computer journal**. ISSN 0010-4620, 1460-2067. 41:8 (1998) 578–588. doi: 10.1093/comjnl/41.8.578.

FURLANUT, M.; FRANCESCHI, L. - Pharmacology of Ifosfamide. **Oncology**. ISSN 0030-2414. 65:2 (2003) 2–6. doi: 10.1159/000073350.

GANDARA, D. R. *et al.* - Squamous Cell Lung Cancer: From Tumor Genomics to Cancer Therapeutics. **Clinical Cancer Research**. ISSN 1078-0432. 21:10 (2015) 2236–2243. doi:

10.1158/1078-0432.CCR-14-3039.

GARCIA, Gwenalyn; ODAIMI, Marcel - Systemic Combination Chemotherapy in Elderly Pancreatic Cancer: a Review. **Journal of Gastrointestinal Cancer**. ISSN 1941-6628. 48:2 (2017) 121–128. doi: 10.1007/s12029-017-9930-0.

GAYVERT, Kaitlyn M. *et al.* - A Computational Approach for Identifying Synergistic Drug Combinations. **PLOS Computational Biology**. ISSN 1553-7358. 13:1 (2017) e1005308. doi: 10.1371/journal.pcbi.1005308.

GHIRMAI, Senait *et al.* - Synthesis and radioiodination of some daunorubicin and doxorubicin derivatives. **Carbohydrate Research**. ISSN 00086215. 340:1 (2005) 15–24. doi: 10.1016/j.carres.2004.10.014.

GILVARY, Coryandar; DRY, Jonathan R.; ELEMENTO, Olivier - Multi-task learning predicts drug combination synergy in cells and in the clinic. **bioRxiv**. (2019) 576017. doi: 10.1101/576017.

GINSBURG, Ophira M. *et al.* - The global cancer epidemic: opportunities for Canada in low- and middle-income countries. **CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne**. ISSN 1488-2329. 184:15 (2012) 1699–704. doi: 10.1503/cmaj.111131.

GRIMM, D. *et al.* - Diagnostic and Therapeutic Use of Membrane Proteins in Cancer Cells. **Current Medicinal Chemistry**. ISSN 09298673. (2011). doi: 10.2174/092986711794088344.

GUPTA, Piyush B. *et al.* - Phenotypic Plasticity: Driver of Cancer Initiation, Progression, and Therapy Resistance. **Cell Stem Cell**. ISSN 19345909. 24:1 (2019) 65–78. doi: 10.1016/j.stem.2018.11.011.

HAN, Xiao; WANG, Junyun; SUN, Yingli - Circulating Tumor DNA as Biomarkers for Cancer Detection. **Genomics, Proteomics & Bioinformatics**. ISSN 1672-0229. 15:2 (2017) 59–72. doi: 10.1016/J.GPB.2016.12.004.

HANAHAN, Douglas; WEINBERG, Robert A. - The hallmarks of cancer. **Cell**. ISSN 0092-8674. 100:1 (2000) 57–70. doi: 10.1016/S0092-8674(00)81683-9.

HANAHAN, Douglas; WEINBERG, Robert A. - Hallmarks of cancer: the next generation. **Cell**. ISSN 1097-4172. 144:5 (2011) 646–74. doi: 10.1016/j.cell.2011.02.013.

HARNED, Theresa M.; GAYNON, Paul S. - Treating refractory leukemias in childhood, role of clofarabine. **Therapeutics and clinical risk management**. ISSN 1176-6336. 4:2 (2008) 327–36. doi: 10.2147/tcrm.s2941.

HARRIS, Timothy J. R.; MCCORMICK, Frank - The molecular pathology of cancer. **Nature Reviews Clinical Oncology**. ISSN 1759-4774. 7:5 (2010) 251–265. doi: 10.1038/nrclinonc.2010.41.

HE, Liye *et al.* - Methods for High-throughput Drug Combination Screening and Synergy Scoring. Humana Press, New York, NY, 2018. ISBN 9781493974931. p. 351–398.

HEYMACH, John *et al.* - Clinical Cancer Advances 2018: Annual Report on Progress Against Cancer From the American Society of Clinical Oncology. **Journal of clinical oncology : official journal of the American Society of Clinical Oncology**. ISSN 1527-7755. 36:10 (2018) 1020–1044. doi: 10.1200/JCO.2017.77.0446.

HOCHREITER, Sepp; SCHMIDHUBER, Jürgen - Long Short-Term Memory. **Neural Computation**. ISSN 0899-7667. 9:8 (1997) 1735–1780. doi: 10.1162/neco.1997.9.8.1735.

HOLBECK, Susan L. *et al.* - The National Cancer Institute ALMANAC: A Comprehensive Screening Resource for the Detection of Anticancer Drug Pairs with Enhanced Therapeutic Activity. **Cancer Research**. ISSN 0008-5472. 77:13 (2017) 3564–3576. doi: 10.1158/0008-5472.CAN-17-0489.

HOMENDA, Władysław; PEDRYCZ, Witold - **Pattern recognition : a quality of data perspective**. Wiley Publishing, Inc., 2018. ISBN 9781119302827.

HORIZON 2020 - Work Programme 2020 - Health, demographic change and wellbeing. [On Line]. Available at WWW:<URL: https://ec.europa.eu/programmes/horizon2020/en/h2020-section/health-demographic-change-and-wellbeing>

HU, Quanyin *et al.* - Recent advances of cocktail chemotherapy by combination drug delivery systems. **Advanced Drug Delivery Reviews**. ISSN 0169-409X. 98 (2016) 19–34. doi: 10.1016/J.ADDR.2015.10.022.

HUANG, Hui-Ling - Propensity Scores for Prediction and Characterization of Bioluminescent Proteins from Sequences. **PLoS ONE**. ISSN 1932-6203. 9:5 (2014) e97158. doi: 10.1371/journal.pone.0097158.

HUMPHREY, William; DALKE, Andrew; SCHULTEN, Klaus - VMD: Visual molecular dynamics. **Journal of Molecular Graphics**. ISSN 02637855. (1996). doi: 10.1016/0263-7855(96)00018-5.

IBRAHIM, Nuha *et al.* - Molecular targeted therapies for cancer: Sorafenib monotherapy and its combination with other therapies (Review). **Oncology Reports**. ISSN 1021335X. 27:5 (2012) 1303–1311. doi: 10.3892/or.2012.1675.

IGNEY, Frederik H.; KRAMMER, Peter H. - Death and anti-death: tumour resistance to apoptosis. **Nature Reviews Cancer**. ISSN 1474-175X. 2:4 (2002) 277–288. doi: 10.1038/nrc776.

JAHAN, Selim - **Human Development Report 2016 - Human Development for Everyone**. The United Nations Development Program, 2016

JAIN, Anil K. - Data clustering: 50 years beyond K-means. **Pattern Recognition Letters**. . ISSN 0167-8655. 31:8 (2010) 651–666. doi: 10.1016/J.PATREC.2009.09.011.

JAMES, Gareth *et al.* - Tree-Based Methods. In **An Introduction to Statistical Learning**. ISBN 9781461471387. p. 303–335.

JANIZEK, Joseph D.; CELIK, Safiye; LEE, Su-In - Explainable machine learning prediction of

synergistic drug combinations for precision cancer medicine. **bioRxiv**. (2018) 331769. doi: 10.1101/331769.

JEMAL, A. *et al.* - Global Patterns of Cancer Incidence and Mortality Rates and Trends. **Cancer Epidemiology Biomarkers & Prevention**. ISSN 1538-7755. 19:8 (2010) 1893–1907. doi: 10.1158/1055-9965.epi-10-0437.

JONES, S.; THORNTON, J. M. - Protein-protein interactions: a review of protein dimer structures. **Progress in biophysics and molecular biology**. ISSN 0079-6107. 63:1 (1995) 31–65.

JORDAN, M. I.; MITCHELL, T. M. - Machine learning: Trends, perspectives, and prospects. **Science**. ISSN 0036-8075. 349:6245 (2015) 255–260. doi: 10.1126/SCIENCE.AAA8415.

KALARI, Krishna R. *et al.* - PANOPLY: Omics-Guided Drug Prioritization Method Tailored to an Individual Patient. **JCO Clinical Cancer Informatics**. ISSN 2473-4276. 2:2 (2018) 1–11. doi: 10.1200/CCI.18.00012.

KALEMKERIAN, Gregory P. - Combination chemotherapy for relapsed small-cell lung cancer. **The Lancet Oncology**. ISSN 14702045. 17:8 (2016) 1033–1035. doi: 10.1016/S1470-2045(16)30160-7.

KAMPEN, Kim R. - Membrane proteins: The key players of a cancer cell. **Journal of Membrane Biology**. ISSN 00222631. (2011). doi: 10.1007/s00232-011-9381-7.

KELLOGG, Glen E.; SCARSDALE, J. Neel; FORNARI, Frank A. - Identification and hydropathic characterization of structural features affecting sequence specificity for doxorubicin intercalation into DNA double-stranded polynucleotides. **Nucleic Acids Research**. ISSN 03051048. 26:20 (1998) 4721–4732. doi: 10.1093/nar/26.20.4721.

KENAKIN, Terry P. - Drug Antagonism. In **Pharmacology in Drug Discovery**. Elsevier, 2012. ISBN 9780123848567. p. 51–80.

KIM, Minseon; OH, Ilhwan; AHN, Jaegyoon - An Improved Method for Prediction of Cancer Prognosis by Network Learning. **Genes**. ISSN 2073-4425. 9:10 (2018) 478. doi: 10.3390/genes9100478.

KIM, Sunghwan *et al.* - PubChem 2019 update: improved access to chemical data. **Nucleic Acids Research**. ISSN 0305-1048. 47:D1 (2019) D1102–D1109. doi: 10.1093/nar/gky1033.

KNOX, R. J. *et al.* - Mechanism of cytotoxicity of anticancer platinum drugs: evidence that cis-diamminedichloroplatinum(II) and cis-diammine-(1,1-cyclobutanedicarboxylato)platinum(II) differ only in the kinetics of their interaction with DNA. **Cancer research**. ISSN 0008-5472. 46:4 Pt 2 (1986) 1972–9.

KOBAYASHI, S. *et al.* - Singly-linked catenation and knotting of cisplatin-DNA adduct by DNA topoisomerase I. **Nucleic acids symposium series**. ISSN 02613166. 29 (1993) 137–138.

KOIZUMI, Fumiaki *et al.* - Synergistic interaction between the EGFR tyrosine kinase inhibitor gefitinib ('Iressa') and the DNA topoisomerase I inhibitor CPT-11 (irinotecan) in human colorectal cancer cells. **International Journal of Cancer**. ISSN 00207136. 108:3 (2004) 464–472. doi: 10.1002/ijc.11539.

KOUVARIS, Kostas *et al.* - How Evolution Learns to Generalise: Principles of under-fitting, over-fitting and induction in the evolution of developmental organisation. **arXiv**. (2015). doi: http://arxiv.org/abs/1508.06854

KUMMAR, Shivaani *et al.* - Drug development in oncology: classical cytotoxics and molecularly targeted agents. **British Journal of Clinical Pharmacology**. ISSN 0306-5251. 62:1 (2006) 15–26. doi: 10.1111/j.1365-2125.2006.02713.x.

KYTE, Jack; DOOLITTLE, Russell F. - A simple method for displaying the hydropathic character of a protein. **Journal of Molecular Biology**. ISSN 00222836. 1982). doi: 10.1016/0022-2836(82)90515-0.

LANDIS, J. Richard; KOCH, Gary G. - The Measurement of Observer Agreement for Categorical Data. **Biometrics**. ISSN 0006341X. 33:1 (1977) 159. doi: 10.2307/2529310.

LEAL, Luis G. *et al.* - Identification of disease-associated loci using machine learning for genotype and network data integration. **Bioinformatics**. ISSN 1367-4803. (2019). doi: 10.1093/bioinformatics/btz310.

LI, Fei *et al.* - Crystal structures of translocator protein (TSPO) and mutant mimic of a human polymorphism. **Science**. ISSN 10959203. (2015). doi: 10.1126/science.1260590.

LI, Jing *et al.* - Brief Introduction of Back Propagation (BP) Neural Network Algorithm and Its Improvement. In **Advances in Computer Science and Information Engineering**. Springer, Berlin, Heidelberg, 2012. ISBN 9783642302237. p. 553–558.

LIN, Jianhua - Divergence measures based on the Shannon entropy. **IEEE Transactions on Information Theory**. ISSN 00189448. 37:1 (1991) 145–151. doi: 10.1109/18.61115.

LINS, Laurence; THOMAS, Annick; BRASSEUR, Robert - Analysis of accessible surface of residues in proteins. **Protein science : a publication of the Protein Society**. ISSN 0961-8368. 12:7 (2003) 1406–17. doi: 10.1110/ps.0304803.

LIU, Rong; JIANG, Wenchao; ZHOU, Yanhong - Identifying protein–protein interaction sites in transient complexes with temperature factor, sequence profile and accessible surface area. **Amino Acids**. ISSN 0939-4451. 38:1 (2010) 263–270. doi: 10.1007/s00726-009-0245-8.

LOEBER, Rachel *et al.* - Cross-linking of the DNA repair protein O6-alkylguanine DNA alkyltransferase to DNA in the presence of antitumor nitrogen mustards. **Chemical Research in Toxicology**. ISSN 0893228X. 21:4 (2008) 787–795. doi: 10.1021/tx7004508.

LOEWE, S.; MUISCHNEK, H. - Über Kombinationswirkungen. **Archiv für Experimentelle**

**Pathologie und Pharmakologie**. ISSN 00281298. 114 (1926) 313–26. doi: 10.1007/BF01952257.

LOMIZE, Andrei L. *et al.* - The role of hydrophobic interactions in positioning of peripheral proteins in membranes. **BMC Structural Biology**. ISSN 14726807. (2007). doi: 10.1186/1472-6807-7-44.

LONG, Nguyen *et al.* - High-Throughput Omics and Statistical Learning Integration for the Discovery and Validation of Novel Diagnostic Signatures in Colorectal Cancer. **International Journal of Molecular Sciences**. ISSN 1422-0067. 20:2 (2019) 296. doi: 10.3390/ijms20020296.

LOTFI-JAM, Kerryann *et al.* - Nonpharmacologic Strategies for Managing Common Chemotherapy Adverse Effects: A Systematic Review. **Journal of Clinical Oncology**. ISSN 0732-183X. 26:34 (2008) 5618–5629. doi: 10.1200/JCO.2007.15.9053.

LOWY, Douglas *et al.* - Cancer moonshot countdown. **Nature Biotechnology**. ISSN 1087-0156. 34:6 (2016) 596–599. doi: 10.1038/nbt.3616.

MARTINS, J. M. *et al.* - Solvent-accessible surface area: How well can be applied to hot-spot detection? **Proteins: Structure, Function and Bioinformatics**. ISSN 08873585. 82:3 (2014). doi: 10.1002/prot.24413.

MATOS-FILIPE, Pedro *et al.* - MENSADB: A Thorough Structural Analysis of Membrane Protein Dimers. **arXiv**. (2019). doi: http://arxiv.org/abs/1902.02321

MAYER, Deborah K.; NASSO, Shelley Fuld - Cancer moonshot: What it means for patients. **Clinical Journal of Oncology Nursing**. ISSN 1538067X. 21:2 (2017) 141–142. doi: 10.1188/17.CJON.141-142.

MAYER, L. D.; JANOFF, A. S. - Optimizing Combination Chemotherapy by Controlling Drug Ratios. **Molecular Interventions**. ISSN 1534-0384. 7:4 (2007) 216–223. doi: 10.1124/mi.7.4.8.

MCKINNEY, Wes - Data Structures for Statistical Computing in Python. In **PROC. OF THE 9th PYTHON IN SCIENCE CONF (SCIPY 2010)** [On Line]. Available at WWW:<URL: https://conference.scipy.org/proceedings/scipy2010/mckinney.html>

**Medicine data: European public assessment reports (EPAR) for veterinary medicines** - [On Line]. European Union, 2018, atual. 2018. [Accessed 1 apr. 2019]. Available at WWW:<URL:https://data.europa.eu/euodp/data/dataset/veterinary-medicines-published-by-the-ema>.

MEESTER, Ronald - Simulation of biological evolution and the NFL theorems. **Biology & philosophy**. ISSN 0169-3867. 24:4 (2009) 461–472. doi: 10.1007/s10539-008-9134-x.

MENDEN, Michael P. *et al.* - Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. **Nature Communications**. ISSN 2041-1723. 10:1 (2019) 2674. doi: 10.1038/s41467-019-09799-2.

MIAO, Lei; HUANG, Leaf - Exploring the Tumor Microenvironment with Nanoparticles. In

**Cancer treatment and research**. 166. p. 193–226.

MOBADERSANY, Pooya *et al.* - Predicting cancer outcomes from histology and genomics using convolutional networks. **Proceedings of the National Academy of Sciences**. ISSN 0027-8424. 115:13 (2018) E2970--E2979. doi: 10.1073/pnas.1717139115.

MOHRI, Mehryar; ROSTAMIZADEH, Afshin; TALWALKAR, Ameet - **Foundations of Machine Learning**. MIT Press, 2012. ISBN 9780262018258.

MOREIRA, Irina S. *et al.* - SpotOn: High Accuracy Identification of Protein-Protein Interface Hot-Spots. **Scientific Reports**. ISSN 20452322. (2017). doi: 10.1038/s41598-017-08321-2.

MORGAN, H. L. - The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. **Journal of Chemical Documentation**. ISSN 0021-9576. 5:2 (1965) 107–113. doi: 10.1021/c160017a018.

MORIWAKI, Hirotomo *et al.* - Mordred: a molecular descriptor calculator. **Journal of Cheminformatics**. ISSN 1758-2946. 10:1 (2018) 4. doi: 10.1186/s13321-018-0258-y.

MUNOS, Bernard - Can open-source R&D reinvigorate drug research? **Nature Reviews Drug Discovery**. ISSN 14741776. 5:9 (2006) 723–729. doi: 10.1038/nrd2131.

MUNOS, Bernard - Lessons from 60 years of pharmaceutical innovation. **Nature Reviews Drug Discovery**. ISSN 14741776. 8:12 (2009) 959–968. doi: 10.1038/nrd2961.

MURRAY, Peter - Organisational learning, competencies, and firm performance: empirical observations. **The Learning Organization**. ISSN 0969-6474. 10:5 (2003) 305–316. doi: 10.1108/09696470310486656.

MUTHUKADAN, Baiju - Selenium with Python [On Line]. (2011). Available at: WWW:<URL: https://selenium-python.readthedocs.io/>

NEUBIG, Richard R. - International Union of Pharmacology Committee on Receptor Nomenclature and Drug Classification. XXXVIII. Update on Terms and Symbols in Quantitative Pharmacology. **Pharmacological Reviews**. ISSN 0031-6997. 55:4 (2003) 597–606. doi: 10.1124/pr.55.4.4.

PASTUR-ROMAY, Lucas *et al.* - Deep Artificial Neural Networks and Neuromorphic Chips for Big Data Analysis: Pharmaceutical and Bioinformatics Applications. **International Journal of Molecular Sciences**. ISSN 1422-0067. 17:8 (2016) 1313. doi: 10.3390/ijms17081313.

PATEL, Jyoti D. *et al.* - Clinical Cancer Advances 2013: Annual Report on Progress Against Cancer From the American Society of Clinical Oncology. **Journal of Clinical Oncology**. ISSN 0732-183X. 32:2 (2014) 129–160. doi: 10.1200/JCO.2013.53.7076.

PAUTIER, Patricia *et al.* - Trabectedin in combination with doxorubicin for first-line treatment of advanced uterine or soft-tissue leiomyosarcoma (LMS-02): a non-randomised, multicentre, phase 2 trial. **The Lancet Oncology**. ISSN 1470-2045. 16:4 (2015) 457–464. doi: 10.1016/S1470-

2045(15)70070-7.

PEDREGOSA, Fabian *et al.* - Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**. ISSN 15337928. 12:Oct (2011) 2825–2830.

PELLINO, Gianluca *et al.* - Noninvasive Biomarkers of Colorectal Cancer: Role in Diagnosis and Personalised Treatment Perspectives. **Gastroenterology Research and Practice**. ISSN 1687-6121. 2018 (2018) 1–21. doi: 10.1155/2018/2397863.

PON, Julia R.; MARRA, Marco A. - Driver and Passenger Mutations in Cancer. **Annual Review of Pathology: Mechanisms of Disease**. ISSN 1553-4006. 10:1 (2015) 25–50. doi: 10.1146/annurev-pathol-012414-040312.

PRAGER, Gerald W. *et al.* - Global cancer control: responding to the growing burden, rising costs and inequalities in access. **ESMO Open**. 3:2 (2018) e000285. doi: 10.1136/esmoopen-2017-000285.

PRETO, António J. *et al.* - Structural Characterization of Membrane Protein Dimers. In **Protein Supersecondary Structures**. Humana Press, New York, NY, 2019. ISBN 9781493991617. p. 403–436.

PREUER, Kristina *et al.* - DeepSynergy: predicting anti-cancer drug synergy with Deep Learning. **Bioinformatics**. ISSN 1367-4803. 34:9 (2018) 1538–1546. doi: 10.1093/bioinformatics/btx806.

PUTIN, Evgeny *et al.* - Reinforced Adversarial Neural Computer for de Novo Molecular Design. **Journal of Chemical Information and Modeling**. ISSN 15205142. 58:6 (2018) 1194–1204. doi: 10.1021/acs.jcim.7b00690.

RAI, Kanti R. *et al.* - Fludarabine Compared with Chlorambucil as Primary Therapy for Chronic Lymphocytic Leukemia. **New England Journal of Medicine**. ISSN 0028-4793. 343:24 (2000) 1750–1757. doi: 10.1056/NEJM200012143432402.

RAMAZZOTTI, Daniele *et al.* - Learning mutational graphs of individual tumour evolution from single-cell and multi-region sequencing data. **BMC Bioinformatics**. ISSN 1471-2105. 20:1 (2019) 210. doi: 10.1186/s12859-019-2795-4.

REITER, Sebastian *et al.* - On the metabolism of allopurinol. **Biochemical Pharmacology**. ISSN 00062952. 32:14 (1983) 2167–2174. doi: 10.1016/0006-2952(83)90222-8.

RITCHIE, Marylyn D. *et al.* - Methods of integrating data to uncover genotype–phenotype interactions. **Nature Reviews Genetics**. ISSN 1471-0056. 16:2 (2015) 85–97. doi: 10.1038/nrg3868.

ROGERS, David; HAHN, Mathew - Extended-Connectivity Fingerprints. **Journal of Chemical Information and Modeling**. ISSN 1549-9596. 50:5 (2010) 742–754. doi: 10.1021/ci100050t.

ROKACH, Lior; MAIMON, Oded - Clustering Methods. In **Data Mining and Knowledge**

**Discovery Handbook**. New York : Springer-Verlag, 2005. p. 321–352.

ROSENBLATT, Frank - The perceptron: A probabilistic model for information storage and organization in the brain. **Psychological Review**. . ISSN 0033295X. 65:6 (1958) 386–408. doi: 10.1037/h0042519. ISBN 9780387254654

ROST, Burkhard; SANDER, Chris - Conservation and prediction of solvent accessibility in protein families. **Proteins: Structure, Function, and Genetics**. ISSN 0887-3585. 20:3 (1994) 216–226. doi: 10.1002/prot.340200303.

RUSSELL, Stuart J.; NORVIG, Peter - **Artificial Intelligence: A Modern Approach**. 3. ed. New Jersey : Pearson Education, Inc., 2010. ISBN 9780136067382.

SAIDIJAM, Massoud; AZIZPOUR, Sonia; PATCHING, Simon G. - Comprehensive analysis of the numbers, lengths and amino acid compositions of transmembrane helices in prokaryotic, eukaryotic and viral integral membrane proteins of high-resolution structure. **Journal of Biomolecular Structure and Dynamics**. ISSN 0739-1102. 36:2 (2018) 443–464. doi: 10.1080/07391102.2017.1285725.

SCANNELL, Jack W. *et al.* - Diagnosing the decline in pharmaceutical R&D efficiency. **Nature Reviews Drug Discovery**. ISSN 14741776. 11:3 (2012) 191–200. doi: 10.1038/nrd3681.

SCHNEIDER, Gisbert - Automating drug discovery. **Nature Reviews Drug Discovery**. . ISSN 1474-1776. 17:2 (2017) 97–113. doi: 10.1038/nrd.2017.232.

SCHWABACHER, Mark - A Survey of Data-Driven Prognostics. In **Infotech@Aerospace**. Reston, Virigina : American Institute of Aeronautics and Astronautics, 26 Sep. 2005. ISBN 9781624100697

SHALEV-SHWARTZ, Shai; BEN-DAVID, Shai - **Understanding Machine Learning**. Cambridge : Cambridge University Press, 2014. ISBN 9781107298019.

SHARMA, Sreenath V.; HABER, Daniel A.; SETTLEMAN, Jeff - Cell line-based platforms to evaluate the therapeutic efficacy of candidate anticancer agents. **Nature Reviews Cancer**. ISSN 1474-175X. 10:4 (2010) 241–253. doi: 10.1038/nrc2820.

SHOEMAKER, Robert H. - The NCI60 human tumour cell line anticancer drug screen. **Nature Reviews Cancer**. ISSN 1474-175X. 6:10 (2006) 813–823. doi: 10.1038/nrc1951.

SIDOROV, Pavel *et al.* - Predicting synergism of cancer drug combinations using NCI-ALMANAC data. **Frontiers in Chemistry**. ISSN 2296-2646. 7 (2019) 509. doi: 10.3389/FCHEM.2019.00509.

SIMÕES, Sérgio *et al.* - On the formulation of pH-sensitive liposomes with long circulation times. **Advanced Drug Delivery Reviews**. ISSN 0169-409X. 56:7 (2004) 947–965. doi: 10.1016/J.ADDR.2003.10.038.

SINGH, Harpreet; RANA, Prashant Singh; SINGH, Urvinder - Prediction of drug synergy in cancer using ensemble-based machine learning techniques. **Modern Physics Letters B**. ISSN 0217-

9849. 32:11 (2018) 1850132. doi: 10.1142/S0217984918501324.

SINHA, Sandeep; VOHORA, Divya - Drug Discovery and Development. In **Pharmaceutical Medicine and Translational Clinical Research**. Elsevier, 2018. ISBN 9780128021033. p. 19–32.

SIVA, Nayanah - 1000 Genomes project. **Nature Biotechnology**. ISSN 1087-0156. 26:3 (2008) 256–256. doi: 10.1038/nbt0308-256b.

STEPHENS, Zachary D. *et al.* - Big Data: Astronomical or Genomical? **PLOS Biology**. ISSN 1545-7885. 13:7 (2015) e1002195. doi: 10.1371/journal.pbio.1002195.

SUN, X.; VILAR, S.; TATONETTI, N. P. - High-Throughput Methods for Combinatorial Drug Discovery. **Science Translational Medicine**. ISSN 1946-6234. 5:205 (2013) 205rv1-205rv1. doi: 10.1126/scitranslmed.3006667.

TALLARIDA, Ronald J. - Revisiting the Isobole and Related Quantitative Methods for Assessing Drug Synergism. **Journal of Pharmacology and Experimental Therapeutics**. ISSN 1521-0103. 342:1 (2012) 2–8. doi: 10.1124/jpet.112.193474.

TANG, Yong *et al.* - Biomarkers for early diagnosis, prognosis, prediction, and recurrence monitoring of non-small cell lung cancer. **OncoTargets and Therapy**. ISSN 11786930. 10 (2017) 4527–4534. doi: 10.2147/OTT.S142149.

TOPORCOV, Tatiana N.; WÜ, Victor; FILHO, Nsch - Epidemiological science and cancer control. **Clinics**. 73 (2018). doi: 10.6061/clinics/2018/e627s.

TOUW, Wouter G. *et al.* - A series of PDB-related databanks for everyday needs. **Nucleic Acids Research**. ISSN 13624962. 2015). doi: 10.1093/nar/gku1028.

ULMSCHNEIDER, Martin B.; SANSOM, Mark S. P. - Amino acid distributions in integral membrane protein structures. **Biochimica et Biophysica Acta - Biomembranes**. ISSN 00052736. (2001). doi: 10.1016/S0005-2736(01)00299-1.

USHIJIMA, Toshikazu; ASADA, Kiyoshi - Aberrant DNA methylation in contrast with mutations. **Cancer Science**. ISSN 13479032. 101:2 (2010) 300–305. doi: 10.1111/j.1349-7006.2009.01434.x.

VERHEUL, Henk M. W. *et al.* - Clinical implications of drug resistance. In **Drug Resistance In The Treatment of Cancer**. Cambridge : Cambridge University Press, 1998. p. 199–231.

VINEIS, Paolo; WILD, Christopher P. - Global cancer patterns: causes and prevention. **The Lancet**. ISSN 01406736. 383:9916 (2014) 549–557. doi: 10.1016/S0140-6736(13)62224-2.

WAGSTAFF, K. *et al.* - Constrained k-means clustering with background knowledge. In **Eighteenth International Conference on Machine Learning**

WANG, Guoli; DUNBRACK, Roland L. - PISCES: a protein sequence culling server. **Bioinformatics**. ISSN 1367-4803. 19:12 (2003) 1589–1591. doi: 10.1093/bioinformatics/btg224.

WEBB, Benjamin; SALI, Andrej - Comparative Protein Structure Modeling Using MODELLER. In

**Current Protocols in Bioinformatics**. Hoboken, NJ, USA : John Wiley & Sons, Inc., 2016 v. 54. p. 5.6.1-5.6.37.

WEBSTER, Rachel M. - Combination therapies in oncology. **Nature Reviews Drug Discovery**. ISSN 1474-1776. 15:2 (2016) 81–82. doi: 10.1038/nrd.2016.3.

WEININGER, David - SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. **Journal of Chemical Information and Modeling**. ISSN 1549-9596. 28:1 (1988) 31–36. doi: 10.1021/ci00057a005.

WHITE, Stephen H. - Biophysical dissection of membrane proteins. **Nature**. ISSN 00280836. 459:7245 (2009) 344–346. doi: 10.1038/nature08142.

WOOD, Derrick E. *et al.* - A machine learning approach for somatic mutation discovery. **Science translational medicine**. ISSN 1946-6242. 10:457 (2018) eaar7939. doi: 10.1126/scitranslmed.aar7939.

WOODS, Derek; TURCHI, John J. - Chemotherapy induced DNA damage response: convergence of drugs and pathways. **Cancer biology & therapy**. ISSN 1555-8576. 14:5 (2013) 379–89. doi: 10.4161/cbt.23761.

WORLD HEALTH ORGANISATION - **Guidelines for ATC Classification and DDD Assignment 2019**. (2019) 22. ed. ISBN 17264898.

XUE-WEN CHEN; XIAOTONG LIN - Big Data Deep Learning: Challenges and Perspectives. **IEEE Access**. ISSN 2169-3536. 2 (2014) 514–525. doi: 10.1109/ACCESS.2014.2325029.

YADAV, Bhagwan *et al.* - Searching for Drug Synergy in Complex Dose-Response Landscapes Using an Interaction Potency Model. **Computational and structural biotechnology journal**. ISSN 2001-0370. 13 (2015) 504–13. doi: 10.1016/j.csbj.2015.09.001.

YAN, Changhui *et al.* - Characterization of Protein–Protein Interfaces. **The Protein Journal**. ISSN 1572-3887. 27:1 (2008) 59–70. doi: 10.1007/s10930-007-9108-x.

YANG, Wanjuan *et al.* - Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. **Nucleic Acids Research**. ISSN 0305-1048. 41:D1 (2012) D955–D961. doi: 10.1093/nar/gks1111.

YANG, Yang *et al.* - Improvement in the Anticancer Activity of 6-Mercaptopurine <i>via</i> Combination with Bismuth(III). **CHEMICAL & PHARMACEUTICAL BULLETIN**. ISSN 0009-2363. 64:11 (2016) 1539–1545. doi: 10.1248/cpb.c15-00949.

YOUNG, Steven R. *et al.* - Optimizing deep learning hyper-parameters through an evolutionary algorithm. In **Proceedings of the Workshop on Machine Learning in High-Performance Computing Environments - MLHPC '15**. New York, New York, USA : ACM Press, 2015. ISBN 9781450340069

YU, Hui *et al.* - Architectures and accuracy of artificial neural network for disease classification

from omics data. **BMC Genomics**. ISSN 1471-2164. 20:1 (2019) 167. doi: 10.1186/s12864-019-5546-z.

YUAN, Guo-Xun; HO, Chia-Hua; LIN, Chih-Jen - Recent Advances of Large-Scale Linear Classification. **Proceedings of the IEEE**. ISSN 0018-9219. 100:9 (2012) 2584–2603. doi: 10.1109/JPROC.2012.2188013.

YUAN, Salina; NORGARD, Robert J.; STANGER, Ben Z. - Cellular Plasticity in Cancer. **Cancer Discovery**. ISSN 2159-8274. 9:7 (2019) 837–851. doi: 10.1158/2159-8290.CD-19-0015.

YUAN, Yuchen *et al.* - DeepGene: an advanced cancer type classifier based on deep learning and somatic point mutations. **BMC Bioinformatics**. ISSN 1471-2105. 17:S17 (2016) 476. doi: 10.1186/s12859-016-1334-9.

ZAHREDDINE, Hiba; BORDEN, Katherine L. B. - Mechanisms and insights into drug resistance in cancer. **Frontiers in pharmacology**. ISSN 1663-9812. 4 (2013) 28. doi: 10.3389/fphar.2013.00028.

ZALUPSKI, M.; BAKER, L. H. - Ifosamide. **JNCI Journal of the National Cancer Institute**. ISSN 0027-8874. 80:8 (1988) 556–566. doi: 10.1093/jnci/80.8.556.

ZEWAIL-FOOTE, Maha *et al.* - The inefficiency of incisions of ecteinascidin 743-DNA adducts by the UvrABC nuclease and the unique structural feature of the DNA adducts can be used to explain the repair-dependent toxicities of this antitumor agent. **Chemistry and Biology**. ISSN 10745521. 8:11 (2001) 1033–1049. doi: 10.1016/S1074-5521(01)00071-0.

ZHANG, Junjun *et al.* - International cancer genome consortium data portal-a one-stop shop for cancer genomics data. **Database**. ISSN 17580463. (2011) bar026. doi: 10.1093/database/bar026.

ZHANG, Lu *et al.* - From machine learning to deep learning: progress in machine intelligence for rational drug discovery. **Drug Discovery Today**. ISSN 1359-6446. 22:11 (2017) 1680–1685. doi: 10.1016/J.DRUDIS.2017.08.010.

ZHANG, Qingchen *et al.* - A survey on deep learning for big data. **Information Fusion**. ISSN 1566-2535. 42 (2018) 146–157. doi: 10.1016/J.INFFUS.2017.10.006.

ZHANG, Shao Qing *et al.* - The membrane- and soluble-protein helix-helix interactome: Similar geometry via different interactions. **Structure**. ISSN 18784186. (2015). doi: 10.1016/j.str.2015.01.009.

ZHANG, Yao *et al.* - Experimental and Computational Evaluation of Forces Directing the Association of Transmembrane Helices. **Journal of the American Chemical Society**. ISSN 0002-7863. 131:32 (2009) 11341–11343. doi: 10.1021/ja904625b.

ZHANG, Zhe; WITHAM, Shawn; ALEXOV, Emil - On the role of electrostatics in protein-protein interactions. **Physical biology**. ISSN 1478-3975. 8:3 (2011) 035001. doi: 10.1088/1478-3975/8/3/035001.

ZHOU, Jianbiao - The role of the tumor microenvironment in hematological malignancies and implication for therapy. **Frontiers in Bioscience**. ISSN 10939946. 10:1–3 (2005) 1581. doi: 10.2741/1642.

ZHOU, Zhi-Hua - **Ensemble Methods**. Chapman and Hall/CRC, 2012. ISBN 9781439830055.

ZIEGEL, Eric *et al.* - **Numerical Recipes: The Art of Scientific Computing**. 3. ed. Cambridge : Cambridge University Press, 2007. ISBN 9780521880688.

# Appendix 1

## *"Elbow" Method*

The elbow method was designed to find the appropriate number of clusters depending on the data in analysis. By applying this method, one assures that he picks the number of clusters that reduce the variation between the samples within the cluster (represented by the Within-cluster Sum of Squares – WCSS (Appendix Equations 1 and 2).

$$WCSS = \sum_{k=1}^{K} \sum_{i=1}^{n_k} z_{ik} \left(x_i - \overline{x}_k\right)^2$$

Appendix Equation 1

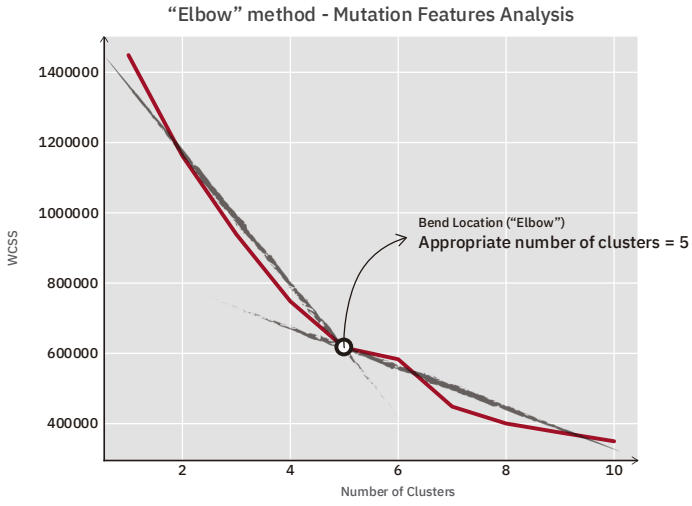
where $K$ is the number of clusters, $n_k$ is the number of elements in the $k^{th}$ cluster, $z_{ik}$ is an indicator function

$$z_{ik} = \begin{cases} 1, & x_i \in cluster\ k \\ 0, & x_i \notin cluster\ k \end{cases}$$

Appendix Equation 2

Method:

1. Compute the $K$-means clustering algorithm for different number of $K$ (usually between 1 and 10);
2. For each $K$, calculate the total WCSS;
3. Plot the curve of $K$ against WCSS;
4. The location of a bend (an "elbow") in the curve is generally considered the appropriate number of clusters (Appendix Figure 1).




Appendix Figure 1: Explanation of the "Elbow" method.

# Annexes

Supplementary Table 1: Summary of all available combination chemotherapeutics, their active substances and the types of cancer suited for treatment. This data was manually curated from the United States' National Cancer Institute Portal (*Comprehensive Cancer Information - National Cancer Institute*, 2019) and from the Cancer UK Portal (*Cancer Research UK*, 2019).

| Drug Designation [a] | Active Substances | Cancer Types Suited for Treatment |
|---|---|---|
| ABVD | Doxorubicin Hydrochloride + Bleomycin + Vinblastine Sulphate + Dacarbazine | Hodgkin Lymphoma |
| ABVE | Doxorubicin Hydrochloride + Bleomycin + Vinblastine Sulphate + Etoposide Phosphate | Hodgkin Lymphoma |
| ABVE-PC | Doxorubicin Hydrochloride + Bleomycin + Vinblastine Sulphate + Etoposide Phosphate + Prednisone + Cyclophosphamide | Hodgkin Lymphoma |
| AC | Doxorubicin Hydrochloride + Cyclophosphamide | Breast Cancer |
| AC-T | Doxorubicin Hydrochloride + Paclitaxel + Cyclophosphamide | Breast Cancer |
| ACE | Adriamycin + Cyclophosphamide + Etoposide | Lung Cancer |
| ADE | Cytarabine (Ara-C) + Daunorubicin Hydrochloride + Etoposide Phosphate | Myeloproliferative Neoplasms<br>Leukaemia |
| BEACOPP | Doxorubicin Hydrochloride + Bleomycin + Vincristine Sulphate + Etoposide Phosphate + Procarbazine Hydrochloride + Cyclophosphamide + Prednisone | Hodgkin Lymphoma |
| BEAM | Carmustine + Etoposide + Cytarabine + Melphalan | Hodgkin Lymphoma<br>Non-Hodgkin Lymphoma |
| BEP | Bleomycin + Cisplatin + Etoposide Phosphate | Ovarian Cancer<br>Testicular Cancer |
| BUMEL | Busulfan + Melphalan Hydrochloride | Neuroblastoma |
| CAF | Doxorubicin Hydrochloride + Fluoroacil + Cyclophosphamide | Breast Cancer |
| CAPOX | Capecitabine + Oxaliplatin | Colon and Rectal |
| CARBO MV | Carboplatin + Methotrexate + Vinblastine | Bladder Cancer |
| CARBOPLATIN-ETOPOSIDE | Carboplatin + Etoposide | Small Cell Lung Cancer<br>Small Cell Bladder Cancer<br>Small Cell Cervical Cancer |
| CARBOPLATIN-TAXOL | Paclitaxel + Carboplatin | Lung Cancer<br>Ovarian Cancer |
| CAV | Cyclophosphamide + Doxorubicin + Vincristine | Lung Cancer |
| CAVE | Cyclophosphamide + Doxorubicin + Vincristine + Etoposide | Small Cell Lung Cancer |
| CEM | Melphalan Hydrochloride + Carboplatin + Etoposide Phosphate | Neuroblastoma |
| CEV | Vincristine Sulphate + Carboplatin + Etoposide Phosphate | Retinoblastoma |
| CHLVPP | Chlorambucil + Vinblastine + Procarbazine + Prednisolone | Hodgkin Lymphoma |
| CHOP | Doxorubicin Hydrochloride + Vincristine Sulphate + Prednisone + Cyclophosphamide | Non-Hodgkin Lymphoma<br>Leukaemia |
| CISPLATIN-FLUOROURACIL-TRASTUZUMAB | Cisplatin + Fluorouracil + Trastuzumab | Gastric Cancer |

| Drug Designation [a] | Active Substances | Cancer Types Suited for Treatment |
|---|---|---|
| CISPLATIN-TEYSUNO | Cisplatin + Teysuno | Gastric Cancer |
| CMF | Methotrexate + Fluoroacil + Cyclophosphamide | Breast Cancer |
| CMV | Cisplatin + Methotrexate + Vinblastine | Bladder Cancer |
| COPDAC | Dacarbazine + Vincristine Sulphate + Prednisone + Cyclophosphamide | Hodgkin Lymphoma |
| COPP | Procarbazine Hydrochloride + Vincristine Sulphate + Prednisone + Cyclophosphamide | Hodgkin Lymphoma<br>Non-Hodgkin Lymphoma |
| COPP-ABV | Procarbazine Hydrochloride + Vincristine Sulphate + Doxorubicin Hydrochloride + Bleomycin + Prednisone + Cyclophosphamide + Vinblastine Sulphate | Hodgkin Lymphoma |
| CTD | Cyclophosphamide + Thalidomide + Dexamethasone | Myeloma |
| CVP | Vincristine Sulphate + Prednisone + Cyclophosphamide | Non-Hodgkin Lymphoma<br>Leukaemia |
| CX | Cisplatin + Capecitabine | Various |
| DHAP | Dexamethasone + Cytarabine + Cisplatin | Non-Hodgkin Lymphoma<br>Hodgkin Lymphoma |
| DOXIFOS | Doxorubicin + Ifosfamide | Soft Tissue Sarcoma |
| EC | Epirubicin + Cyclophosphamide | Breast Cancer |
| ECARBOX | Epirubicin + Carboplatin + Capecitabine | Gastro Oesophageal Cancer |
| ECF | Epirubicin + Cisplatin + Fluorouracil | Gastro Oesophageal Cancer |
| ECX | Epirubicin + Cisplatin + Capecitabine | Gastro Oesophageal Cancer |
| EOF | Epirubicin + Oxaliplatin + Fluorouracil | Gastro Oesophageal Cancer |
| EOX | Epirubicin + Oxaliplatin + Capecitabine | Gastro Oesophageal Cancer |
| EPOCH | Etoposide Phosphate + Vincristine Sulphate + Doxorubicin Hydrochloride + Prednisone + Cyclophosphamide | Non-Hodgkin Lymphoma |
| ESHAP | Etoposide + Methylprednisolone + Cytarabine + Cisplatin | Hodgkin Lymphoma<br>Non-Hodgkin Lymphoma<br>Myeloma |
| FAD | Fludarabine + Doxorubicin + Dexamethasone | Low Grade Lymphoma |
| FCR | Fludarabine + Cyclophosphamide + Rituximab | Chronic Lymphocytic Leukaemia |
| FEC | Fluorouracil + Epirubicin Hydrochloride + Cyclophosphamide | Breast Cancer |
| FEC-T | Fluorouracil + Epirubicin Hydrochloride + Cyclophosphamide + Docetaxel | Breast Cancer |
| FLUOROURACIL (5FU)-CISPLATIN | Cisplatin + Fluorouracil | Anal Cancer<br>Head and Neck Cancer<br>Oesophageal Cancer |
| FLUOROURACIL (5FU)-MITOMYCIN C | Fluorouracil + Mitomycin C | Anal Cancer |
| FMD | Fludarabine + Mitoxantrone + Dexamethasone | Non-Hodgkin Lymphoma |
| FOLFIRI | Leucovorin Calcium + Fluorouracil + Irinotecan Hydrochloride | Colon and Rectal Cancer |
| FOLFIRI-BEVACIZUMAB | Leucovorin Calcium + Fluorouracil + Irinotecan Hydrochloride + Bevacizumab | Colon and Rectal Cancer |
| FOLFIRI-CETUXIMAB | Leucovorin Calcium + Fluoroacil + Irinotecan Hydrochloride + Cetuximab | Colon and Rectal Cancer |
| FOLFIRINOX | Leucovorin Calcium + Fluorouracil + Irinotecan Hydrochloride + Oxaliplatin | Pancreatic Cancer |

| Drug Designation [a] | Active Substances | Cancer Types Suited for Treatment |
|---|---|---|
| FOLFOX | Leucovorin Calcium + Fluorouracil + Oxaliplatin | Colon and Rectal Cancer |
| GEMCAP | Gemcitabine Hydrochloride + Capecitabine | Pancreatic Cancer |
| GEMCARBO | Gemcitabine Hydrochloride + Carboplatin | Non-Small Cell Lung Cancer<br>Bladder Cancer<br>Advanced Breast Cancer<br>Ovarian Cancer |
| GEMCITABINE-CISPLATIN | Gemcitabine Hydrochloride + Cisplatin | Bladder Cancer<br>Cervical Cancer<br>Lung Cancer<br>Malignant Mesothelioma<br>Ovarian Cancer<br>Pancreatic Cancer |
| GEMCITABINE-OXALIPLATIN | Gemcitabine Hydrochloride + Oxaliplatin | Pancreatic Cancer |
| GEMTAXOL | Gemcitabine Hydrochloride + Paclitaxel | Breast Cancer<br>Bladder Cancer |
| HYPER-CVAD | Cyclophosphamide + Vincristine Sulphate + Doxorubicin Hydrochloride + Dexamethasone + Methotrexate (optional) + Cytarabine (optional) | Non-Hodgkin Lymphoma |
| HYPER-CVAD (ALL) | Cyclophosphamide + Vincristine Sulphate + Doxorubicin Hydrochloride + Dexamethasone + Methotrexate (optional) + Cytarabine (optional) | Leukaemia |
| ICE | Ifosfamide + Carboplatin + Etoposide Phosphate | Hodgkin Lymphoma<br>Non-Hodgkin Lymphoma |
| JEB | Carboplatin + Etoposide Phosphate + Bleomycin | Ovarian Cancer<br>Testicular Cancer |
| MIC | Mitomycin + Ifosfamide + Cisplatin | Non-Small Cell Lung Cancer<br>Oesophageal Cancer |
| MMM | Mitoxantrone + Mitomycin C + Methotrexate | Breast Cancer |
| MOPP | Mechlorethamine Hydrochloride + Vincristine Sulphate + Procarbazine Hydrochloride + Prednisone | Hodgkin Lymphoma |
| MPT | Melphalan + Prednisolone + Thalidomide | Myeloma |
| MVAC | Methotrexate + Vinblastine Sulphate + Doxorubicin Hydrochloride + Cisplatin | Bladder Cancer |
| MVP | Mitomycin + Vinblastine Sulphate + Cisplatin | Lung Cancer<br>Mesothelioma<br>Breast Cancer |
| OEPA | Vincristine Sulphate + Etoposide Phosphate + Prednisone + Doxorubicin Hydrochloride | Hodgkin Lymphoma |
| OFF | Oxaliplatin + Fluoroacil + Leucovorin Calcium | Pancreatic Cancer |
| OPPA | Vincristine Sulphate + Procarbazine Hydrochloride + Prednisone + Doxorubicin Hydrochloride | Hodgkin Lymphoma |
| PAD | Bortezomib + Doxorubicin Hydrochloride + Dexamethasone | Multiple Myeloma and Other Plasma Cell Neoplasms |
| PC | Paclitaxel + Carboplatin | Ovarian Cancer<br>Cervical Cancer<br>Small Cell Lung Cancer |

| Drug Designation [a] | Active Substances | Cancer Types Suited for Treatment |
|---|---|---|
| PCV | Procarbazine Hydrochloride + Lomustine + Vincristine Sulphate | Brain Tumours |
| PE | Etiposide + Cisplatin | Small Cell Lung Cancer<br>Germ Cell Cancers<br>Small Cell of the Neck of the Womb |
| PEB | Cisplatin + Etoposide Phosphate + Bleomycin | Ovarian Cancer<br>Testicular Cancer |
| PMITCEBO | Prednisolone + Mitoxantrone + Cyclophosphamide + Etiposide + Bleomycin + Vincristine | Non-Hodgkin Lymphoma |
| POMB/ACE | Cisplatin + Vincristine Sulphate + Methotrexate + Bleomycin + Actinomycin + Cyclophosphamide + Etoposide | Testicular Cancer |
| R-CHOP | Rituximab + Cyclophosphamide + Doxorubicin Hydrochloride + Vincristine Sulphate + Prednisone | Non-Hodgkin Lymphoma |
| R-CVP | Rituximab + Cyclophosphamide + Vincristine Sulphate + Prednisone | Non-Hodgkin Lymphoma |
| R-DHAP | Rituximab + Dexamethasone + Cytarabine + Cisplatin | Non-Hodgkin Lymphoma |
| R-EPOCH | Rituximab + Etoposide Phosphate + Prednisone + Vincristine Sulphate + Cyclophosphamide + Doxorubicin Hydrochloride | Non-Hodgkin Lymphoma |
| R-ESHAP | Rituximab + Etoposide + Methylprednisolone + Cytarabine + Cisplatin | Lymphoma |
| R-GCVP | Rituximab + Gemcitabine + Cyclophosphamide + Vincristine Sulphate + Prednisolone | Hodgkin Lymphoma |
| R-ICE | Rituximab + Ifosfamide + Carboplatin + Etoposide Phosphate | Non-Hodgkin Lymphoma |
| STANFORD V | Mechlorethamine Hydrochloride + Doxorubicin Hydrochloride + Vinblastine Sulphate + Vincristine Sulphate + Bleomycin + Etoposide Phosphate + Prednisone | Hodgkin Lymphoma |
| TAC | Docetaxel + Doxorubicin Hydrochloride + Cyclophosphamide | Breast Cancer |
| TIP | Paclitaxel + Cisplatin + Ifosfamide | Testicular Cancer |
| TPF | Docetaxel + Cisplatin + Fluorouracil | Head and Neck<br>Stomach (Gastric) Cancer |
| VAC | Vincristine Sulphate + Dactinomycin + Cyclophosphamide | Ovarian Cancer<br>Soft Tissue Sarcoma |
| VAD | Vincristine Sulphate + Doxorubicin Hydrochloride + Dexamethasone | Myeloma |
| VAI | Vincristine Sulphate + Actinomycin + Ifosfamide | Ewing's sarcoma |
| VAMP | Vincristine Sulphate + Doxorubicin Hydrochloride + Methotrexate + Prednisone | Hodgkin Lymphoma |
| VEIP | Vinblastine Sulphate + Ifosfamide + Cisplatin | Ovarian Cancer<br>Testicular Cancer |
| VIDE | Vincristine Sulphate + Ifosfamide + Doxorubicin Hydrochloride + Etoposide Phosphate | Ewing's Sarcoma |
| VIP | Etoposide + Ifosfamide + Cisplatin | Testicular Cancer |
| XELIRI | Capecitabine + Irinotecan Hydrochloride | Colon and Rectal<br>Oesophageal<br>Stomach (Gastric) Cancer |
| XELOX | Capecitabine + Oxaliplatin | Colon and Rectal |

[a], Drug designation is represented in most of the cases by the first capital letter of each active substance (2nd column) in the combination.

Supplementary Table 2: DNN evaluation results per drug.

| Name | ATC Class | acc | auc | f-score | MSE | rec | prec |
|---|---|---|---|---|---|---|---|
| Methotrexate | Folic acid analogues | 0,7695 | 0,5009 | 0,8693 | 0,2305 | 0,9880 | 0,7760 |
| Busulfan | Alkyl sulfonates | 0,7619 | 0,5843 | 0,8556 | 0,2381 | 0,9479 | 0,7797 |
| Thioguanine | Purine analogues | 0,7665 | 0,5271 | 0,8655 | 0,2335 | 0,9897 | 0,7689 |
| Mercaptopurine | Purine analogues | 0,6907 | 0,5194 | 0,8119 | 0,3093 | 0,9631 | 0,7017 |
| Mechlorethamine hydrochloride | Nitrogen mustard analogues | 0,8128 | 0,5606 | 0,8934 | 0,1872 | 0,9805 | 0,8206 |
| Allopurinol | Preparations inhibiting uric acid production | 0,7568 | 0,5052 | 0,8603 | 0,2432 | 0,9751 | 0,7697 |
| Dactinomycin | Actinomycines | 0,6713 | 0,6666 | 0,7054 | 0,3287 | 0,7391 | 0,6746 |
| Chlorambucil | Nitrogen mustard analogues | 0,7344 | 0,5640 | 0,8384 | 0,2656 | 0,9708 | 0,7378 |
| Thiotepa | Ethylene imines | 0,8125 | 0,5403 | 0,8944 | 0,1875 | 0,9863 | 0,8182 |
| Melphalan | Nitrogen mustard analogues | 0,7787 | 0,5206 | 0,8738 | 0,2213 | 0,9836 | 0,7860 |
| Triethylenemelamine | Non-classifyed/Experimental | 0,6458 | 0,6230 | 0,7190 | 0,3542 | 0,8056 | 0,6493 |
| Altretamine | Other antineoplastic agents | 0,8902 | 0,5000 | 0,9419 | 0,1098 | 1,0000 | 0,8902 |
| Quinacrine hydrochloride | Non-classifyed/Experimental | 0,6954 | 0,5713 | 0,8043 | 0,3046 | 0,9298 | 0,7087 |
| Aminolevulinic acid hydrochloride | Sensitizers used in photodynamic/radiation therapy | 0,7706 | 0,5095 | 0,8693 | 0,2294 | 0,9852 | 0,7778 |
| Fluorouracil | Pyrimidine analogues | 0,7255 | 0,5000 | 0,8409 | 0,2745 | 1,0000 | 0,7255 |
| Plicamycin | Other cytotoxic antibiotics | 0,6361 | 0,6209 | 0,5465 | 0,3639 | 0,5056 | 0,5948 |
| Pipobroman | Other alkylating agents | 0,7129 | 0,6370 | 0,8013 | 0,2871 | 0,8918 | 0,7275 |
| Cyclophosphamide | Nitrogen mustard analogues | 0,7788 | 0,4923 | 0,8756 | 0,2212 | 0,9847 | 0,7883 |
| Mitomycin | Other cytotoxic antibiotics | 0,7218 | 0,6286 | 0,8141 | 0,2782 | 0,9310 | 0,7232 |
| Floxuridine | Pyrimidine analogues | 0,7697 | 0,5284 | 0,8673 | 0,2303 | 0,9853 | 0,7746 |
| Hydroxyurea | Other antineoplastic agents | 0,7132 | 0,5346 | 0,8255 | 0,2868 | 0,9459 | 0,7322 |
| Uracil mustard | Nitrosoureas | 0,8021 | 0,5000 | 0,8902 | 0,1979 | 1,0000 | 0,8021 |
| Mitotane | Other antineoplastic agents | 0,6516 | 0,5178 | 0,7805 | 0,3484 | 0,9412 | 0,6667 |
| Dacarbazine | Other alkylating agents | 0,8788 | 0,5000 | 0,9355 | 0,1212 | 1,0000 | 0,8788 |
| Methoxsalen | Psoralens for systemic use | 0,8392 | 0,5064 | 0,9123 | 0,1608 | 0,9924 | 0,8442 |

| Name | ATC Class | acc | auc | f-score | MSE | rec | prec |
|---|---|---|---|---|---|---|---|
| Vinblastine sulfate | Vinca alkaloids and analogues | 0,6816 | 0,6693 | 0,7316 | 0,3184 | 0,7830 | 0,6866 |
| Cytarabine hydrochloride | Pyrimidine analogues | 0,7235 | 0,5693 | 0,8282 | 0,2765 | 0,9462 | 0,7364 |
| Thalidomide | Other immunosuppressants | 0,7057 | 0,4938 | 0,8261 | 0,2943 | 0,9631 | 0,7232 |
| Vincristine sulfate | Vinca alkaloids and analogues | 0,6136 | 0,6212 | 0,6200 | 0,3864 | 0,6992 | 0,5569 |
| Megestrol acetate | Progestogens | 0,6383 | 0,6061 | 0,7385 | 0,3617 | 0,6667 | 0,8276 |
| Procarbazine hydrochloride | Methylhydrazines | 0,7500 | 0,5460 | 0,8518 | 0,2500 | 0,9701 | 0,7592 |
| Lomustine | Nitrosoureas | 0,7475 | 0,5024 | 0,8530 | 0,2525 | 0,9367 | 0,7831 |
| Daunorubicin hydrochloride | Anthracyclines and related substances | 0,6343 | 0,6205 | 0,5294 | 0,3657 | 0,5714 | 0,4932 |
| Streptozocin | Nitrosoureas | 0,7419 | 0,5465 | 0,8471 | 0,2581 | 0,9965 | 0,7366 |
| Arsenic trioxide | Other antineoplastic agents | 0,6438 | 0,5559 | 0,7554 | 0,3563 | 0,8408 | 0,6857 |
| Azacitidine | Pyrimidine analogues | 0,6692 | 0,6102 | 0,7668 | 0,3308 | 0,9079 | 0,6636 |
| Cladribine | Purine analogues | 0,7395 | 0,6638 | 0,8208 | 0,2605 | 0,8554 | 0,7889 |
| Ifosfamide | Nitrogen mustard analogues | 0,7513 | 0,5239 | 0,8549 | 0,2487 | 0,9716 | 0,7632 |
| 2-Fluoro Ara-A | Purine analogues | 0,7527 | 0,5935 | 0,8475 | 0,2473 | 0,9470 | 0,7669 |
| Cisplatin | Platinum compounds | 0,6357 | 0,6396 | 0,6332 | 0,3643 | 0,5865 | 0,6880 |
| Tretinoin | Other antineoplastic agents | 0,6995 | 0,6281 | 0,7893 | 0,3005 | 0,8655 | 0,7254 |
| Teniposide | Podophyllotoxin derivatives | 0,6601 | 0,6523 | 0,7045 | 0,3399 | 0,7045 | 0,7045 |
| Doxorubicin hydrochloride | Anthracyclines and related substances | 0,7483 | 0,6747 | 0,8304 | 0,2517 | 0,8230 | 0,8378 |
| Bleomycin sulfate | Other cytotoxic antibiotics | 0,6284 | 0,5691 | 0,7295 | 0,3716 | 0,7915 | 0,6766 |
| Paclitaxel | Taxanes | 0,6143 | 0,6143 | 0,6215 | 0,3857 | 0,6333 | 0,6101 |
| Decitabine | Pyrimidine analogues | 0,7010 | 0,6004 | 0,8000 | 0,2990 | 0,8761 | 0,7360 |
| bendamustine hydrochloride | Nitrogen mustard analogues | 0,7045 | 0,5742 | 0,8097 | 0,2955 | 0,8927 | 0,7409 |
| Etoposide | Podophyllotoxin derivatives | 0,7615 | 0,7302 | 0,8243 | 0,2385 | 0,8133 | 0,8356 |
| Dexrazoxane | Detoxifying agents for antineoplastic treatment | 0,6992 | 0,5024 | 0,8213 | 0,3008 | 0,9770 | 0,7083 |
| Tamoxifen citrate | Anti-estrogens | 0,6933 | 0,6720 | 0,7525 | 0,3067 | 0,7917 | 0,7170 |
| Pentostatin | Other antineoplastic agents | 0,7013 | 0,4963 | 0,8233 | 0,2987 | 0,9739 | 0,7131 |

| Name | ATC Class | acc | auc | f-score | MSE | rec | prec |
|---|---|---|---|---|---|---|---|
| Sirolimus | Selective immunosuppressants | 0,7052 | 0,6885 | 0,7574 | 0,2948 | 0,8029 | 0,7167 |
| Carboplatin | Platinum compounds | 0,6443 | 0,5597 | 0,7579 | 0,3557 | 0,8816 | 0,6646 |
| Valrubicin | Anthracyclines and related substances | 0,6096 | 0,6078 | 0,5746 | 0,3904 | 0,5423 | 0,6111 |
| Oxaliplatin | Platinum compounds | 0,6298 | 0,5769 | 0,7262 | 0,3702 | 0,7755 | 0,6828 |
| Mitoxantrone | Anthracyclines and related substances | 0,6522 | 0,5461 | 0,2000 | 0,3478 | 0,1154 | 0,7500 |
| Amifostine | Detoxifying agents for antineoplastic treatment | 0,7273 | 0,5052 | 0,8376 | 0,2727 | 0,9007 | 0,7827 |
| Temozolomide | Other alkylating agents | 0,8000 | 0,5938 | 0,8824 | 0,2000 | 0,9507 | 0,8232 |
| Imiquimod | Antivirals | 0,8179 | 0,5185 | 0,8989 | 0,1821 | 0,9894 | 0,8235 |
| Carmustine | Nitrosoureas | 0,7260 | 0,5669 | 0,8304 | 0,2740 | 0,9425 | 0,7421 |
| Clofarabine | Purine analogues | 0,7382 | 0,6183 | 0,8315 | 0,2618 | 0,9453 | 0,7421 |
| Vinorelbine tartrate | Vinca alkaloids and analogues | 0,6897 | 0,6500 | 0,5424 | 0,3103 | 0,5000 | 0,5926 |
| Topotecan hydrochloride | Other antineoplastic agents | 0,6978 | 0,5708 | 0,8056 | 0,3022 | 0,9147 | 0,7198 |
| Gemcitabine hydrochloride | Pyrimidine analogues | 0,8101 | 0,5222 | 0,8940 | 0,1899 | 0,9961 | 0,8109 |
| Docetaxel | Taxanes | 0,6742 | 0,6742 | 0,7038 | 0,3258 | 0,6742 | 0,7362 |
| 7-Ethyl-10-hydroxycamptothecin | Non-classifyed/Experimental | 0,6856 | 0,6226 | 0,7749 | 0,3144 | 0,8077 | 0,7447 |
| Bortezomib | Other antineoplastic agents | 0,6707 | 0,5595 | 0,7816 | 0,3293 | 0,8608 | 0,7158 |
| Nelarabine | Purine analogues | 0,8182 | 0,6834 | 0,8869 | 0,1818 | 0,9739 | 0,8142 |
| Pemetrexed Disodium | Folic acid analogues | 0,7313 | 0,5018 | 0,8421 | 0,2687 | 0,9412 | 0,7619 |
| Vorinostat | Other antineoplastic agents | 0,7087 | 0,7013 | 0,7491 | 0,2913 | 0,7463 | 0,7519 |
| Estramustine phosphate sodium | Other antineoplastic agents | 0,8027 | 0,6995 | 0,8722 | 0,1973 | 0,9305 | 0,8208 |
| Capecitabine | Pyrimidine analogues | 0,7481 | 0,6346 | 0,8371 | 0,2519 | 0,9598 | 0,7422 |
| Exemestane | Aromatase inhibitors | 0,5517 | 0,5270 | 0,6389 | 0,4483 | 0,6866 | 0,5974 |
| Gefitinib | Protein kinase inhibitors | 0,7528 | 0,6447 | 0,4762 | 0,2472 | 0,3846 | 0,6250 |
| Erlotinib hydrochloride | Protein kinase inhibitors | 0,6831 | 0,6632 | 0,7466 | 0,3169 | 0,8385 | 0,6728 |
| Fulvestrant | Anti-estrogens | 0,6711 | 0,6336 | 0,7487 | 0,3289 | 0,7807 | 0,7192 |
| Anastrozole | Aromatase inhibitors | 0,8043 | 0,5557 | 0,8883 | 0,1957 | 0,9890 | 0,8063 |
| Letrozole | Aromatase inhibitors | 0,6633 | 0,6518 | 0,7179 | 0,3367 | 0,7925 | 0,6563 |
| Celecoxib | Sulfonamides | 0,6765 | 0,6649 | 0,7250 | 0,3235 | 0,7632 | 0,6905 |

| Name | ATC Class | acc | auc | f-score | MSE | rec | prec |
|---|---|---|---|---|---|---|---|
| Zoledronic acid | Bisphosphonates | 0,7827 | 0,5000 | 0,8781 | 0,2173 | 1,0000 | 0,7827 |
| Dasatinib | Protein kinase inhibitors | 0,6327 | 0,6319 | 0,6159 | 0,3673 | 0,6159 | 0,6159 |
| Everolimus | Protein kinase inhibitors | 0,7341 | 0,7319 | 0,7624 | 0,2659 | 0,7476 | 0,7778 |
| Pazopanib hydrochloride | Protein kinase inhibitors | 0,8974 | 0,5000 | 0,0000 | 0,1026 | 0,0000 | 0,0000 |
| Imatinib mesylate | Protein kinase inhibitors | 0,7732 | 0,6416 | 0,8571 | 0,2268 | 0,8919 | 0,8250 |
| Lapatinib ditosylate | Protein kinase inhibitors | 0,7385 | 0,7112 | 0,7958 | 0,2615 | 0,8363 | 0,7590 |
| Nilotinib | Protein kinase inhibitors | 0,7406 | 0,6572 | 0,8265 | 0,2594 | 0,8239 | 0,8291 |
| Sorafenib tosylate | Protein kinase inhibitors | 0,7313 | 0,6178 | 0,8244 | 0,2687 | 0,8970 | 0,7627 |
| Lenalidomide | Other immunosuppressants | 0,7500 | 0,4853 | 0,8571 | 0,2500 | 0,9706 | 0,7674 |
| Ixabepilone | Other cytotoxic antibiotics | 0,6525 | 0,5890 | 0,7500 | 0,3475 | 0,8084 | 0,6995 |
| Raloxifene hydrochloride | Selective estrogen receptor modulators | 0,6400 | 0,5000 | 0,0000 | 0,3600 | 0,0000 | 0,0000 |
| Abiraterone | Other hormone antagonists and related agents | 0,6087 | 0,5798 | 0,4490 | 0,3913 | 0,5000 | 0,4074 |
| Sunitinib (free base) | Protein kinase inhibitors | 0,7658 | 0,5797 | 0,8591 | 0,2342 | 0,9505 | 0,7837 |
| vemurafenib | Protein kinase inhibitors | 0,6703 | 0,6633 | 0,7115 | 0,3297 | 0,7551 | 0,6727 |
| Romidepsin | Other antineoplastic agents | 0,6499 | 0,6191 | 0,4937 | 0,3501 | 0,3959 | 0,6555 |
| pralatrexate | Folic acid analogues | 0,7389 | 0,5000 | 0,8499 | 0,2611 | 1,0000 | 0,7389 |
| Vismodegib | Other antineoplastic agents | 0,6230 | 0,5429 | 0,7356 | 0,3770 | 0,8000 | 0,6809 |
| Crizotinib | Protein kinase inhibitors | 0,6503 | 0,5954 | 0,7444 | 0,3497 | 0,8137 | 0,6860 |
| Axitinib | Protein kinase inhibitors | 0,7045 | 0,5223 | 0,1333 | 0,2955 | 0,0769 | 0,5000 |
| Vandetanib | Protein kinase inhibitors | 0,4643 | 0,5000 | 0,0000 | 0,5357 | 0,0000 | 0,0000 |
| vemurafenib | Protein kinase inhibitors | 0,7358 | 0,6656 | 0,8250 | 0,2642 | 0,7857 | 0,8684 |
| Cabazitaxel | Taxanes | 0,8000 | NaN | NaN | 0,2000 | NaN | 0,0000 |
| Ruxolitinib | Protein kinase inhibitors | 0,5370 | 0,5662 | 0,6575 | 0,4630 | 0,9600 | 0,5000 |

acc = accuracy; ROC AUC = Area Under the Receiver Operating Characteristics Curve; MSE = Mean Squared Error; prec = precision; rec = recall; Gini = Gini coefficient

Supplementary Table 3: DNN evaluation results per cell line.

| Cell Line | Disease | Cancer Type | acc | ROC AUC | f-score | MSE | rec | prec |
|---|---|---|---|---|---|---|---|---|
| HL-60(TB) | Adult acute myeloid leukemia | Blood | 0,6673 | 0,6617 | 0,7034 | 0,3327 | 0,7258 | 0,6824 |
| SR | Anaplastic large cell lymphoma | Blood | 0,6579 | 0,6587 | 0,6537 | 0,3421 | 0,6752 | 0,6336 |
| RPMI-8226 | Plasma cell myeloma | Blood | 0,6785 | 0,6655 | 0,7302 | 0,3215 | 0,7375 | 0,7230 |
| K-562 | Chronic myelogenous leukemia | Blood | 0,6926 | 0,6347 | 0,7779 | 0,3074 | 0,8174 | 0,7420 |
| SF-295 | Glioblastoma | Brain | 0,7599 | 0,6857 | 0,8359 | 0,2401 | 0,8655 | 0,8082 |
| SNB-75 | Glioblastoma | Brain | 0,7194 | 0,6924 | 0,7804 | 0,2806 | 0,8068 | 0,7557 |
| SF-539 | Gliosarcoma | Brain | 0,6908 | 0,6473 | 0,7689 | 0,3092 | 0,7946 | 0,7448 |
| SF-268 | Astrocytoma | Brain | 0,7268 | 0,6876 | 0,7945 | 0,2732 | 0,8371 | 0,7560 |
| U251 | Astrocytoma | Brain | 0,7769 | 0,7174 | 0,8445 | 0,2231 | 0,8671 | 0,8230 |
| T-47D | Invasive ductal carcinoma | Breast | 0,6897 | 0,6521 | 0,7665 | 0,3103 | 0,7665 | 0,7665 |
| HS 578T | Invasive ductal carcinoma | Breast | 0,7069 | 0,6467 | 0,7899 | 0,2931 | 0,8309 | 0,7529 |
| MDA-MB-231/ATCC | Breast adenocarcinoma | Breast | 0,6804 | 0,6283 | 0,7657 | 0,3196 | 0,7905 | 0,7424 |
| 7 MCF | Invasive ductal carcinoma | Breast | 0,7262 | 0,6713 | 0,8034 | 0,2738 | 0,8264 | 0,7817 |
| BT-549 | Invasive ductal carcinoma | Breast | 0,7145 | 0,6736 | 0,7873 | 0,2855 | 0,8029 | 0,7722 |
| MDA-MB-468 | Breast adenocarcinoma | Breast | 0,6759 | 0,6694 | 0,7137 | 0,3241 | 0,7408 | 0,6885 |
| HCT-116 | Colon carcinoma | Colon | 0,7085 | 0,6371 | 0,7958 | 0,2915 | 0,8356 | 0,7596 |
| SW-620 | Colon adenocarcinoma | Colon | 0,7360 | 0,6577 | 0,8206 | 0,2640 | 0,8288 | 0,8127 |
| KM12 | Colon carcinoma | Colon | 0,7500 | 0,5930 | 0,8455 | 0,2500 | 0,8526 | 0,8385 |
| HCT-15 | Colon adenocarcinoma | Colon | 0,7660 | 0,6982 | 0,8398 | 0,2340 | 0,8532 | 0,8269 |
| 786-0 | Renal cell carcinoma | Kidney | 0,6957 | 0,6725 | 0,7609 | 0,3043 | 0,7543 | 0,7675 |
| ACHN | Papillary renal cell carcinoma | Kidney | 0,7031 | 0,6786 | 0,7644 | 0,2969 | 0,7913 | 0,7392 |
| CAKI-1 | Clear cell renal cell carcinoma | Kidney | 0,7518 | 0,6746 | 0,8348 | 0,2482 | 0,8212 | 0,8489 |
| A498 | Renal cell carcinoma | Kidney | 0,7344 | 0,6867 | 0,8065 | 0,2656 | 0,8299 | 0,7843 |

| Cell Line | Disease | Cancer Type | acc | ROC AUC | f-score | MSE | rec | prec |
|---|---|---|---|---|---|---|---|---|
| UO-31 | Renal cell carcinoma | Kidney | 0,7324 | 0,6697 | 0,8109 | 0,2676 | 0,8443 | 0,7800 |
| NCI-H460 | Large cell lung carcinoma | Lung | 0,7295 | 0,7060 | 0,7885 | 0,2705 | 0,7918 | 0,7853 |
| HOP-92 | Non-small cell lung carcinoma | Lung | 0,7488 | 0,6458 | 0,8351 | 0,2512 | 0,8639 | 0,8081 |
| NCI-H322M | Minimally invasive lung adenocarcinoma | Lung | 0,7312 | 0,6673 | 0,8120 | 0,2688 | 0,8223 | 0,8020 |
| HOP-62 | Lung adenocarcinoma | Lung | 0,7385 | 0,7071 | 0,8021 | 0,2615 | 0,8083 | 0,7959 |
| EKVX | Lung adenocarcinoma | Lung | 0,7440 | 0,6749 | 0,8240 | 0,2560 | 0,8309 | 0,8172 |
| NCI-H522 | Lung adenocarcinoma | Lung | 0,7234 | 0,6743 | 0,7978 | 0,2766 | 0,8292 | 0,7688 |
| A549/ATCC | Lung adenocarcinoma | Lung | 0,7321 | 0,6862 | 0,8092 | 0,2679 | 0,7919 | 0,8274 |
| NCI-H23 | Lung adenocarcinoma | Lung | 0,6777 | 0,6446 | 0,7505 | 0,3223 | 0,7946 | 0,7111 |
| IGROV1 | Ovarian endometrioid adenocarcinoma | Ovarian | 0,7228 | 0,6745 | 0,7986 | 0,2772 | 0,8090 | 0,7885 |
| OVCAR-4 | High grade ovarian serous adenocarcinoma | Ovarian | 0,7513 | 0,6600 | 0,8357 | 0,2487 | 0,8404 | 0,8311 |
| OVCAR-3 | High grade ovarian serous adenocarcinoma | Ovarian | 0,7418 | 0,6627 | 0,8244 | 0,2582 | 0,8433 | 0,8065 |
| SK-OV-3 | Ovarian serous cystadenocarcinoma | Ovarian | 0,7197 | 0,6637 | 0,7980 | 0,2803 | 0,8374 | 0,7622 |
| OVCAR-8 | High grade ovarian serous adenocarcinoma | Ovarian | 0,7210 | 0,6622 | 0,8009 | 0,2790 | 0,8268 | 0,7766 |
| DU-145 | Prostate carcinoma | Prostate | 0,7091 | 0,6864 | 0,7687 | 0,2909 | 0,7840 | 0,7540 |
| PC-3 | Prostate carcinoma | Prostate | 0,6849 | 0,6427 | 0,7631 | 0,3151 | 0,7969 | 0,7321 |
| MALME-3M | Melanoma | Skin | 0,7304 | 0,6929 | 0,7991 | 0,2696 | 0,8062 | 0,7921 |
| UACC-62 | Melanoma | Skin | 0,6594 | 0,6431 | 0,7205 | 0,3406 | 0,7147 | 0,7265 |
| SK-MEL-5 | Cutaneous melanoma | Skin | 0,6730 | 0,6493 | 0,7383 | 0,3270 | 0,7604 | 0,7174 |
| SK-MEL-28 | Cutaneous melanoma | Skin | 0,7313 | 0,6506 | 0,8163 | 0,2687 | 0,8501 | 0,7851 |
| UACC-257 | Melanoma | Skin | 0,7136 | 0,6489 | 0,7986 | 0,2864 | 0,8090 | 0,7885 |
| LOX IMVI | Amelanotic melanoma | Skin | 0,7382 | 0,6670 | 0,8206 | 0,2618 | 0,8245 | 0,8167 |

acc = accuracy; ROC AUC = Area Under the Receiver Operating Characteristics Curve; MSE = Mean Squared Error; prec = precision; rec = recall; Gini = Gini coefficient