Master en bioinformatique et modélisation

From predicting drug response in cancer cell lines to personalized oncology

Simon PENELLE

Mémoire présenté en vue de l'obtention du diplôme d'un Master en Bioinformatique et Modélisation

Supervisé par:

Prof. Marianne ROOMAN, Prof. Fabrizio PUCCI

Computational Biology and Bioinformatics Lab

Academic Year 2024-2025

Remerciements **2**
Abstract **4**
Résumé en français **5**
Chapter 1 – Background **6**
1.1 Context and Motivation 6
1.2 Definition of Cancer 7
1.3 Sensitivity/Resistance Mechanisms 11
1.4 Genomics of Drug Sensitivity in Cancer (GDSC) 15
1.5 BRAF 17
1.6 State of the Art 22
Chapter 2 – Objectives **25**
Chapter 3 – Methods **26**
Chapter 4 – Results **33**
4.1 GDSC dataset description 33
4.2 Genetic variations 36
4.3 Structural analysis 38
4.3.1 Alpha Fold 3 predictions 38
4.3.2 Clustering (RMSD- pLDDT) 43
4.3.3 Protein-Protein and Protein-ligand Binding affinity predictions 47
4.3.4 Building a structural predictor for drug-target identification 53
4.3.5 Correlations with IC50 54
4.4 Expression data 56
Chapter 5 – Discussion and Perspectives **58**
5.1 Structural bioinformatics for drug response prediction 59
5.2 Strengths and limitations of the approach 59
5.3 Future directions and perspectives 60
Chapter 6 – Conclusion **61**
References **64**
Appendix A **68**

Remerciements

Je tiens tout d'abord à remercier mes promoteurs et professeurs, Marianne Rooman et Fabrizio Pucci, pour leurs conseils et l'expertise qu'ils m'ont apportés tout au long de ce mémoire. À de nombreuses reprises, ils m'ont aidé à ne pas me perdre dans mes recherches, en me redirigeant vers des approches plus prometteuses, tout en me laissant une grande liberté dans mon travail. Je leur suis également reconnaissant pour l'accueil chaleureux qu'ils m'ont réservé au sein des laboratoires de l'ULB malgré ma demande tardive pour réaliser mon mémoire chez eux.

Je souhaite également remercier les membres du laboratoire 3BIO ainsi que ceux du laboratoire IRIBHM à Erasme, avec qui j'ai eu de nombreux échanges très enrichissants et qui étaient toujours prêts à donner un coup de main si nécessaire. J'ai sincèrement apprécié évoluer dans un environnement composé de personnes passionnées par leur travail et par la bioinformatique en général.

Je tiens tout particulièrement à remercier Benoît, qui m'a encadré tout au long de ce mémoire et qui a toujours pris le temps de répondre à mes questions malgré les responsabilités liées à sa propre thèse. Au début de ce travail, mes connaissances en oncologie étaient limitées, mais il m'a beaucoup aidé à me familiariser avec le sujet. Pouvoir discuter avec quelqu'un d'aussi renseigné que lui m'a permis de mieux appréhender ce sujet fascinant

Enfin, je remercie ma famille et mes amis pour leur soutien incroyable. En particulier, Camille, qui m'a remonté le moral dans les moments difficiles, ainsi que mes parents, qui m'ont encouragé et accompagné tout au long de ces longues années d'études.

Abstract

The main objective of this thesis is to explore whether integrating structural biology features into predictive pipelines can improve our ability to predict cancer cell sensitivity to treatment. Using data from existing genomic datasets like GDSC, this work aims to go one step further by adding protein-level structural information using modern deep learning tools. In this thesis, we focus on the BRAF gene, a well-studied oncogene with clinically relevant mutations such as V600E, to assess whether structural differences between wild-type and mutated forms can explain variations in drug response. To achieve this goal, I developed a computational pipeline combining open-source tools such as Alpha Fold 3 (for structure prediction), PyMOL (for visualization), and Boltz 2 (for estimating ligand–protein binding affinity). A second analysis layer based on gene expression levels was added to allow the contribution of multiple biological dimensions (genomic, transcriptomic, and structural) to the observed variability in drug response. The results of this analysis show a modest correlation of 0.255 between our predictor and experimental data, mainly due to limitations of the deep learning models used and the complex, multidimensional and evolving nature of cancer. Ultimately, this thesis aims to evaluate the performance of deep learning–based structural predictors in a multi-omics drug discovery context, and to determine to what extent these approaches can contribute to identifying, in silico, cancer cell lines that are sensitive or resistant to therapy.

Keywords: Bioinformatics, Structural biology, Personalized oncology, Multi-omics, BRAF mutations, Protein-ligand interaction, GDSC dataset, Deep learning

Résumé en français

L'objectif principal de ce mémoire est d'explorer si l'intégration de caractéristiques issues de la biologie structurale dans des pipelines prédictifs peut améliorer notre capacité à prédire la sensibilité des cellules cancéreuses aux traitements. En s'appuyant sur des jeux de données génomiques existants comme GDSC, ce travail vise à aller un pas plus loin en incorporant des informations structurales au niveau des protéines à l'aide d'outils modernes basés sur l'intelligence artificielle. Dans ce mémoire, nous analysons plus particulièrement le gène BRAF, un oncogène bien étudié présentant des mutations cliniquement pertinentes comme V600E. Le but est d'évaluer si les différences de structure entre les formes sauvage et mutée peuvent expliquer des variations dans la réponse aux médicaments. Pour cela, j'ai développé un pipeline computationnel combinant des outils open source comme Alpha Fold 3 (prédiction structurale), PyMOL (visualisation) et Boltz 2 (affinité protéine-ligand). Une seconde couche d'analyse, basée sur les niveaux d'expression génique, a été ajoutée pour permettre la contribution de multiples dimensions (génomique, transcriptomique et structurelle) à la variabilité de réponse observée. Les résultats de cette analyse montrent une corrélation modeste de 0.255 entre notre prédicteur et les données expérimentales principalement due aux limitations des modèles d'apprentissage profond utilisés mais aussi de l'aspect multidimensionnel et évolutif du cancer. Au final, ce mémoire vise à évaluer la performance des prédicteurs structuraux basés sur le deep learning dans un contexte de découverte de médicaments multi-omique, et à déterminer dans quelle mesure ces approches peuvent contribuer à identifier in silico des lignées cellulaires sensibles ou résistantes aux thérapies.

Mots-clés : Bioinformatique, Biologie structurelle, Cancérologie personnalisée, Multi-omique, Mutations BRAF, Interactions protéine-ligand, GDSC, Apprentissage profond

Chapter 1 – Background

1.1 Context and Motivation

Cancer is one of the most widespread illnesses today with over ~40% of the population expected to have at least one type of cancer in their lifetime [1]. For decades, advances in oncology have drastically reduced the number of lethal cases by carefully crafting and selecting drugs to improve cancer prognosis. Many treatments exist today like immunotherapy where we try to help the immune system recognize and attack cancer cells. Chemotherapy is another one of the most successful approaches to impair the development of cancer lines. Unfortunately chemotherapeutic agents come with the cost of huge side effects resulting from their cytotoxic action. While those therapeutics usually work well for a majority of people, some patients develop resistance to those drugs after a while indicating an adaptation of the cancer to circumvent the drug lethality. These adaptations are primarily driven by single or multiple mutations within cancer cells, leading to enhanced survival, drug resistance, and reduced competition under selective pressure from specific therapies. The prediction of drug response in cancer cell lines has become a key strategy for identifying molecular factors that influence treatment efficacy. Datasets such as Genomics of Drug Sensitivity in Cancer (GDSC) have provided fundamental information to link genomic and transcriptomic features to drug sensitivity. In this master's thesis, we aim to develop a computational framework to predict drug responsiveness by integrating multiple layers of biological information and using state-of-the-art deep learning tools. The goal is to extract and engineer features from genomic and transcriptomic data, while also incorporating protein structural information to enhance the biological interpretability of the model. Beyond modeling drug response in cancer cell lines, we will extend our approach to a more clinically relevant scenario by accounting for tumor clonal heterogeneity, a key challenge in personalized oncology. In this phase, we selected a specific driver mutation and assessed its potential impact on drug sensitivity. This was achieved by

predicting how these mutations alter protein function and, in turn, influence treatment response. The ultimate goal of this work is to move toward a precision oncology framework, where we can computationally predict the most effective drugs or drug combinations for individual patients based on their unique molecular profiles.

1.2 Definition of Cancer

Cancer is one of the most common diseases associated with aging in human populations with 1 over 3 persons estimated to develop cancer in their lifetime [1] . Since its discovery and characterization, many efforts have been deployed worldwide to better understand cancer and find effective treatments to improve patient's prognosis and avoid post-treatment relapse. Cancer exists in different forms in human and model organisms and is defined as "a disease of uncontrolled proliferation by transformed cells subject to evolution by natural selection". [2]
The evolution feature of this definition is defined by changes in gene frequency within a population of cells. Those can involve adaptations leading to phenotypic modifications of the cell that can be favored by the tumor microenvironment. Many of those advantageous adaptations revolve around uptaking resources, exploiting normal cells such as fibroblasts, evading the immune system or recruiting pro-tumor components, producing new blood vessels and parts of the extracellular matrix, creating and surviving acidic conditions. Indeed, adaptations of cancer cells that emerge from evolution by natural selection result in many hallmarks of cancer [figure 1] so updates of the definition of cancer should include the force of natural selection acting on the initiation and progression of the cancer cell populations. [2]

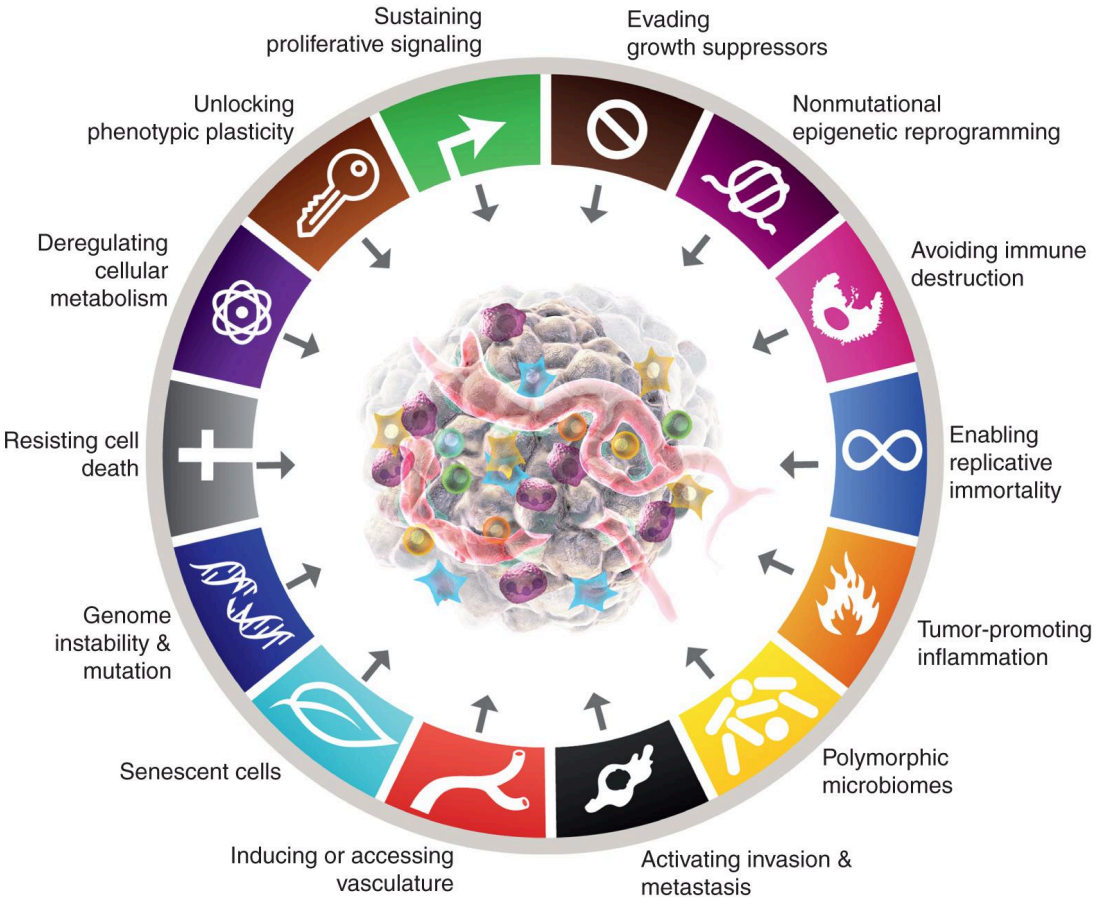


Figure 1: *The hallmarks of cancer, a set of characteristics that cancer cells acquire during their development, which enable them to grow uncontrollably and spread throughout the body (from [3])*

One of the main hallmarks of cancer development relies on genome instability and mutations in the genetic code of healthy cells leading to some advantageous phenotype in terms of better survival and reproduction. Those mutations can be classified according to the type of advantage it gives to tumor cell population. While some mutations have deleterious, or absence of impact on survival or development of tumor cells like passenger mutations [4], others can give an adaptive advantage to the cell compared to surrounding cells in a specific organ leading to a disproportion in clonal population sizes. Such mutations called driver mutations are the main cause of cancer development and can be classified as oncogenes or tumor suppressor genes (TSG). Oncogenes are defined as genes that promote cancer development when abnormally activated, for example BRAF or KRAS genes controlling cell division and proliferation via the RAS/MAPK pathway. TSG's are genes that prevent cancer when

functioning normally. Mutations in those genes can induce a loss-of-function that may lead to the promotion of cancer. Many of those deleterious mutations are classified in driver gene databases for example DriverDB, TCGA's Pan-Cancer Atlas or BoostDM. [5,6,7]

The TCGA publicly funded project aims to characterize and catalogue important cancer-causing biological alterations (for example mutations, DNA methylation or transcriptional profiles) and associate them to 33 cancer types including 10 rare cancers at a large scale using more than 11,000 cases of cancer samples [Figure 2].

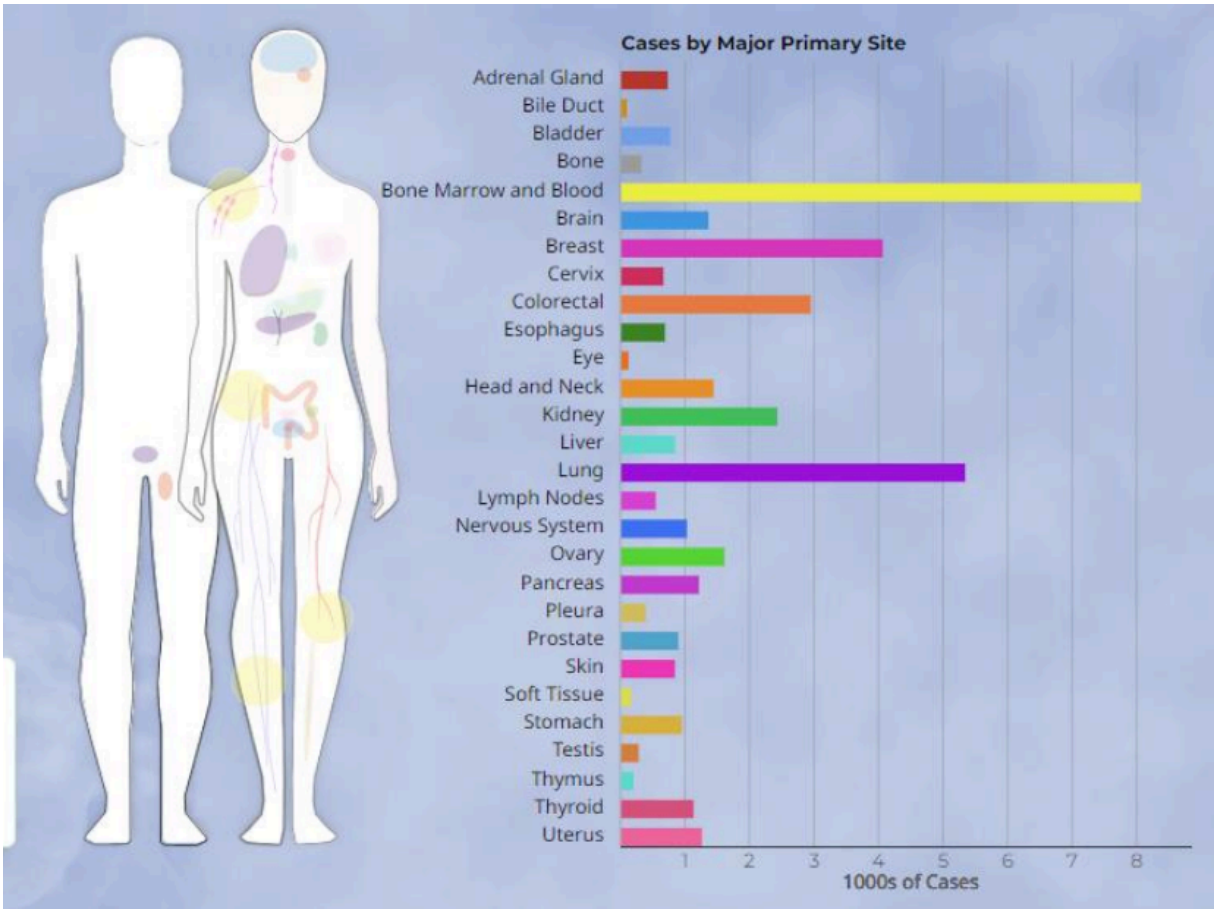


Figure 2: *TCGA classification of cancer cell lines based on tissue of origin, grouping them into biologically relevant cancer types (from [8])*

One of the main paradigms of cancer development relies on clonal expansion, where similar cancer cells (clones) grow and compete for limited resources like oxygen and nutrients. These clones also interact with the tumor microenvironment which includes immune cells, fibroblasts, and can influence how the tumor evolves. Many of those

advantageous mutations revolve around increasing proliferation, high-jacking resources of nearby tissues, resisting drugs or evading the immune system. For example the V600E mutation in the BRAF protein can lead to a dysregulation of the MAPK/ERK pathway resulting in uncontrolled proliferation of tumor cells. This pathway is crucial in the regulation of important cellular processes such as growth, differentiation, and survival. Cells that acquire such mutations will rapidly multiply and outnumber surrounding cells and might acquire more deleterious mutations.

Many innovative treatments are available today to treat cancer, including targeted chemotherapy, immunotherapy, precision surgery, combination therapies, and adjuvant radiotherapies. These innovations have driven a 32% decline in overall cancer death rate between 1991 and 2019, averting 3.5 million deaths [10]. Those great advances in cancer survival rates over the last 50 years are not homogeneous over all cancer types. Some types of cancer have a high survival rate today like testis or breast cancer, while others still painfully hover around less than 20%, for example Lung and Pancreas are the most dangerous types of cancer with a survival rate close to 0% [figure 3]. Those encouraging results can be explained by the presence of cheap and effective screening procedures, disparities in therapeutics development or tumor heterogeneity between different types of cancer.

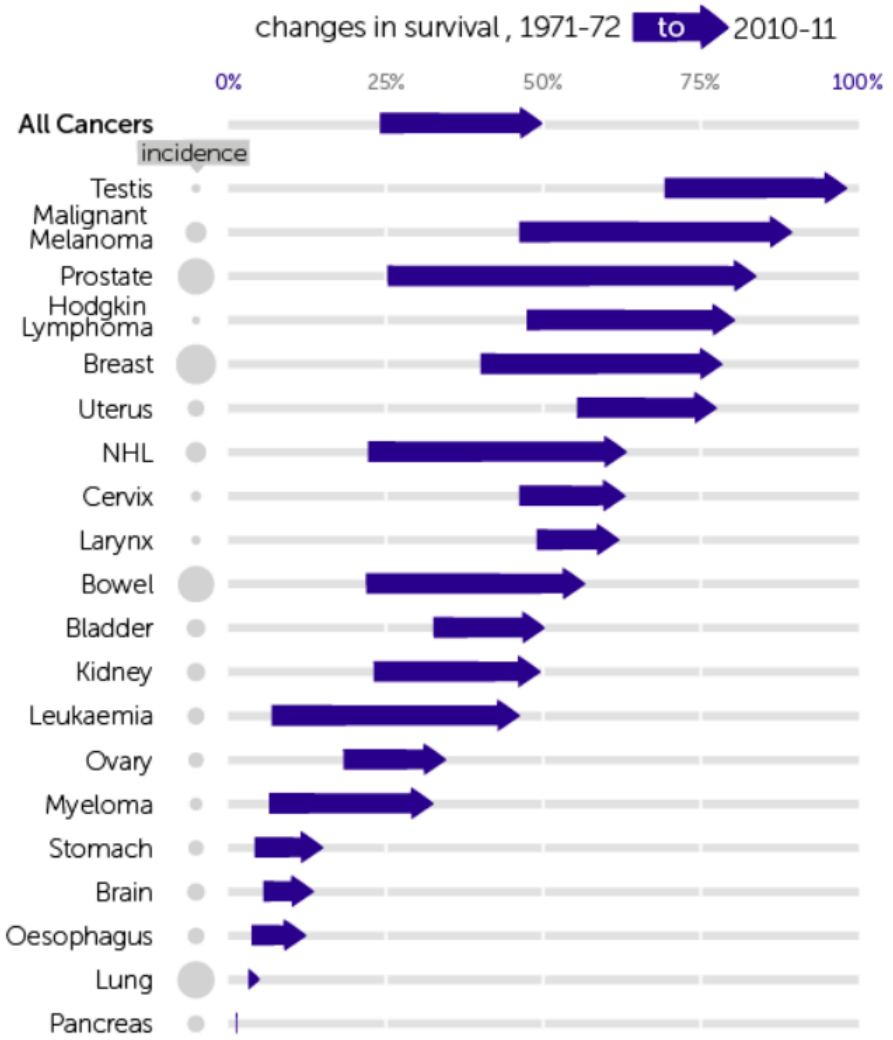


Figure 3: *Cancer survival rate evolution from 1971 to 2011 (from [9])*

1.3 Sensitivity/Resistance Mechanisms

One of the major causes of failure for cancer treatments has been the phenomenon of cancer resistance that can lead to poor prognosis in cancer patients. This phenomenon is a complex and multi aspect mechanism that can arise from genetic, epigenetic, protein expression levels and/or environmental factors. In current times it is still a commonly observed phenomenon in various forms of cancer even with the

recent improvements in cancer treatment and better therapeutics development.

A strategy to avoid resistance has been prescribing multiple drug combinations to patients to overcome a specific resistance mechanism, but cells often become cross-resistant to a variety of cancer drugs. This multi-drug resistance (MDR) phenomenon is a major hurdle in many cases and is the main cause of cancer relapse and cancer related death. Drug resistance can be categorized as intrinsic or primary resistance, meaning that the cell acquired the resistant feature before the first drug dose administration, or acquired resistance that develops in response to receiving the cancer treatment and is the consequence of adaptive modifications in initially sensitive cancer cell populations leading to less effective treatment over time. There are many identified mechanisms that are associated with drug resistance exploiting different genetic or non-genetic factors. Many factors in tumor cells have been considered critical contributors to therapeutic resistance, including acquired mutations of drug targets, amplification of oncogenes in compensatory or bypass pathways, and epigenetic modifications, which can further affect intratumor heterogeneity, tumor cell plasticity, DNA repair, and the susceptibility of tumor cells to cell death pathways. Those mechanisms are not mutually exclusive and might stack to reduce therapeutic responsiveness [11]. Strategies to overcome cancer resistance include biomarker identification for predicting resistance, developing new drugs that target specific resistance mechanisms or combining therapeutics targeting multiple detrimental pathways at the same time.

The next section describes common well studied and characterised mechanisms that might lead to increased drug resistance or sensitivity in cellular populations, they are all illustrated in figure 4. Other mechanisms not listed below are described in the literature but are less common and relevant to this study.

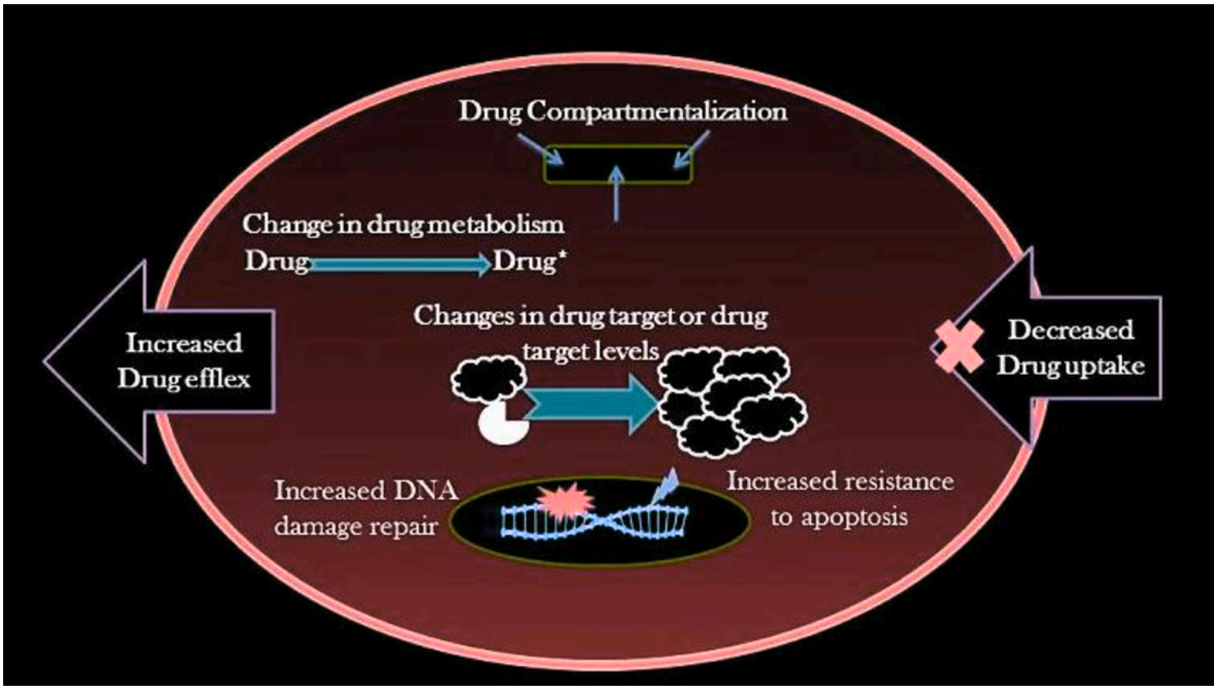


Figure 4: *Possible mechanisms of drug resistance (from [12])*

A. Increased Drug efflux

ABC transporters are important efflux pumps that have a role in reducing toxic concentrations and moving various molecules across cell membranes using the energy from ATP hydrolysis. In some cases they can have a huge role in drug resistance by reducing intracellular drug accumulation by acting as drug efflux facilitators, leading to multi-drug resistance (MDR)

B. Decreased drug uptake

Another way resistance can arise is when the drug is unable to enter the cell because of mechanisms such as expression levels or mutations reducing drug influx. Most of the time this mechanism is present for specific membrane transporters responsible for the influx of certain drugs. For example Methotrexate enters cells by means of the reduced folate carrier, and decreased expression of this protein results in more severe resistance to the drug. [13]

C. Drug compartmentalization

Intracellular organelles are sometimes able to sequester drugs into vesicles like lysosomes. This prevents the drug from interacting with the target site and decreases the drug effectiveness. Such compartments can also metabolically modify the drug before releasing it in the cytosol where the cytotoxic effect may be disturbed or inactivated.

D. Change in drug metabolism

Many drugs are disassembled by enzymes (ex. cytochrome P450s) into their main metabolites before being integrated into the cells. Those mechanisms can inactivate drugs or decrease their prodrug activity by reducing blood concentrations of the active metabolite.

E. Change in drug target or drug target levels

Drug targets are sometimes mutated in resistant cells to be unaffected by the initially active drug and the decreased affinity can lead to resistance. Other strategies are the overexpression of slightly altered targets that saturate drug binding molecules requiring much higher dosing of the drug to have the same biological effect.

F. Increase DNA damage repair

Cancer cells often have modified capabilities when looking at DNA modification and repair mechanisms. This involves up- or down regulation of DNA repair pathways leading to more mutations or facilitating aberrant DNA modifications and better survival chances after therapy induced DNA damage.

G. Increased resistance to apoptosis

Apoptosis is a key cellular regulation mechanism that forces damaged cells to be inactivated and eliminated by the immune system to avoid spreading of damaging molecules or uncontrolled proliferation of the damaged cell. Typically the p53 protein is a major molecular actor in the apoptosis inducing regulation but inactivation or modifications to this protein can lead to resistance to apoptosis despite the cytotoxic damage induced by anticancer drugs.

1.4 Genomics of Drug Sensitivity in Cancer (GDSC)

Over the years of trying to combat and help people getting rid of all forms of complex diseases, especially cancer, researchers in oncology started to develop important bioinformatics tools for cancer research. One of those tools is the publication of enormous public databases containing many important biological information about cancer and patients. This innovation is a major milestone for drug discovery and genetic characterization of different types of cancer related diseases. Such databases are a revolution in the field, they give access to high quality biological data from thousands of labs all over the world. Those resources are still very important today in the field as many researchers try to develop multi-parameter predictors or new drug combinations to improve cancer prognosis. Those databases can contain a huge variety of information about cancer cases and patient biology, for example driver gene databases (e.g. BoostDM) [7], drug databases (e.g. Pubchem) [14], protein structure databases (e.g. PDB, Uniprot) [15,16], metabolic databases (e.g KEGG), but also databases targeting specifically cancer like CCLE, TCGA, or GDSC [18,8,19]. GDSC and CCLE are two of the most complete and popular databases of cancer cell lines containing over 1000 different cancer cell lines and drug interactions. Cancer cell lines are grown in vitro in a controlled environment to use them as an experimental model to test multiple drugs. More complex models include organoids, which are 3D cell cultures that can mimic different tissues, and xenografts, where human tumor cells are implanted into mice to study cancer in a living system. Figure 5 shows an overview of all cancer experimental models that are used in clinical research. Those models can be categorized depending on the scale of the models (at the cellular level, at the organ level, …). New clinical hypotheses are often validated on smaller and simpler models before being tested on larger ones.

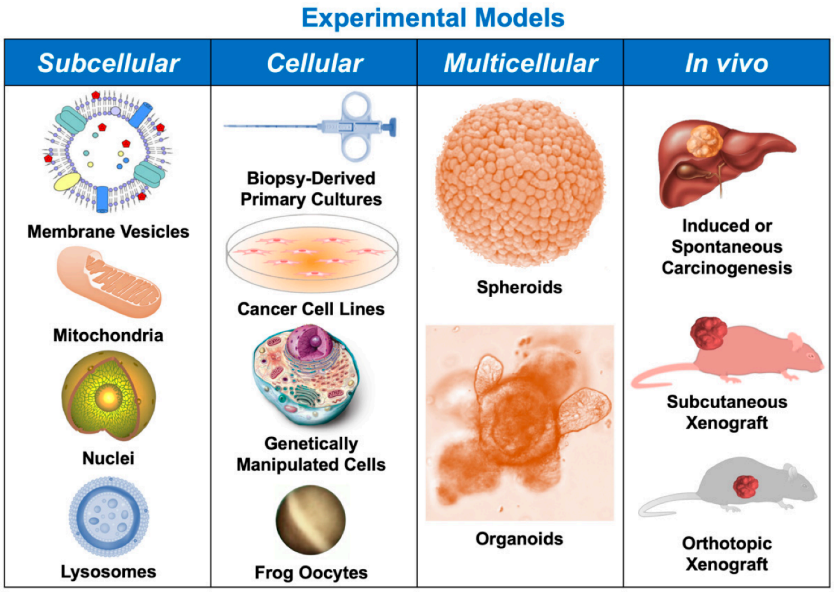


Figure 5: *Classification of all cancer experimental models used in pharmaceutical research (from [20])*

The *'Genomics of Drug Sensitivity in Cancer database'* (GDSC) was first released in 2012 and contained over 700 cancer cell lines, 138 anticancer drugs and over 75000 experiments of drug/cell line interaction. Today it has 624 anticancer drugs, 979 cell lines and more than 500.000 sensitivity values for drug/cell lines pairs. The sensitivity is described in terms of IC50 which is a quantitative measure used in pharmacology to indicate the concentration of a drug needed to inhibit a specific biological process or target by 50%, where lower IC50 indicates higher sensitivity to the drug. The main objective of the GDSC database is to characterize the sensitivity or resistance of all included cell lines to the most common anticancer drugs, using high throughput techniques (complete experimental automatisation) to determine IC50 values for each cell-line/drug interaction, as is illustrated in figure 6. In addition to the sensitivity information, it contains extensively characterized biological information about the cell lines at the genomic (driver mutations, WGS, WES), epigenomic (methylation pattern, chromatin accessibility), transcriptomic, and proteomic level. All available data that can be downloaded directly from on the website's free download portal.

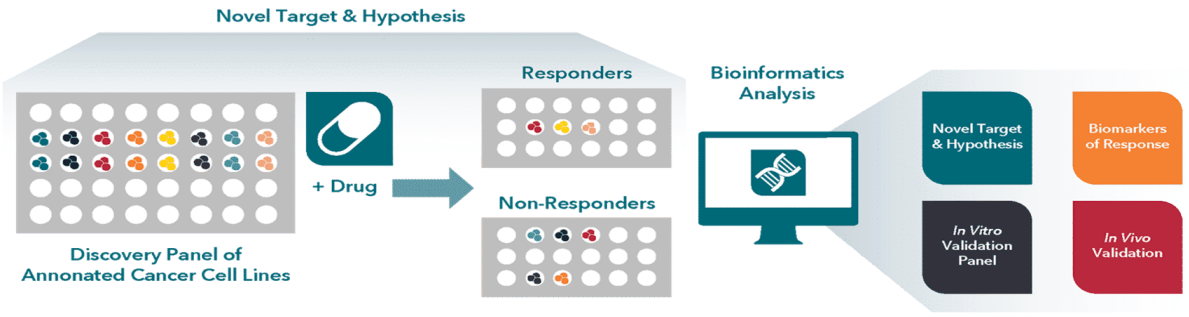


Figure 6: *High-throughput automated screening uses robotic systems to systematically test large numbers of drug–cell line combinations, enabling rapid assessment of drug responses (from [19])*

1.5 BRAF

In this section, we explore one of the most represented and studied oncogenes, *BRAF*, which is mutated in a significant proportion of human cancers. Approximately 50% of melanomas, around 10–20% of colorectal cancers [21,25], and lower percentages for lung and thyroid cancers. *BRAF* encodes a serine/threonine kinase that plays a central role in the RAS/RAF/MEK/ERK pathway, a signaling cascade essential for the regulation of cell proliferation and survival. The *BRAF* gene is located on chromosome 7q34, encoding a protein of 766 amino acids [22]. Uncontrolled activation of the RAS/RAF/MEK/ERK pathway is a common strategy of many cancers giving them the possibility to proliferate in an uncontrolled manner, strongly increasing the probability of new harmful mutations. In the following sections we describe the details of the RAS/RAF/MEK/ERK pathway and how BRAF protein is a key regulator of the proliferation pathway. Then, we explore a classification of different common mutations in *BRAF* and related targets that are found in many cancer cell lines. The last section explores the timeline of different drugs and therapeutic strategies to inhibit the unregulated overexpression of the pathway by targeting BRAF or other adjacent molecular targets.

The RAF family of serine/threonine kinases includes A-RAF, B-RAF, and C-RAF. Those variants share three conserved regions called CR1, CR2, and CR3 [Figure 7]. Those domains regulate the interaction with RAS protein in active form. CR1 contains the RAS-binding domain (RBD) and a cysteine-rich domain (CRD). CR2 is a regulatory domain that interacts with 14-3-3 proteins, forcing the kinase in an autoinhibited state by stabilizing their interaction between the N-terminal and C-terminal domain. CR3 includes the catalytic kinase domain itself, responsible for MEK phosphorylation [27]. In the inactive conformation, CR1 and CR3 domains interact, preventing the kinase activity. Once RAS binds to the RBD, the autoinhibitory contacts are disrupted, releasing 14-3-3 proteins and exposing the dimerization interface crucial for activation. BRAF can then form homodimers or heterodimers with itself or CRAF. The active BRAF dimer binds and activates MEK1 by positioning it near the activation αG-helix of the kinase domain. This facilitates the transfer of phosphate groups from ATP to MEK1. In the activation segment, the DFG (Asp, Phe, Gly) motif positions ATP for catalysis. Figure 8 schematises the molecular dynamic of BRAF activation. On the contrary, if BRAF remains in its autoinhibited conformation, the orientation of MEK1 is incompatible with phosphorylation, and kinase activity is blocked. The dynamic of BRAF is further regulated by ERK that phosphorylates inhibitory sites within RAF proteins, leading to the dissociation from RAS and destabilization of active dimers [23]. To note that BRAF is the only kinase able to phosphorylate MEK. CRAF and, under certain conditions, ARAF can also contribute to MEK activation. On the other hand, phosphorylated MEK1 is the only protein capable of activating ERK1 and ERK2 to ensure precise substrate specificity.

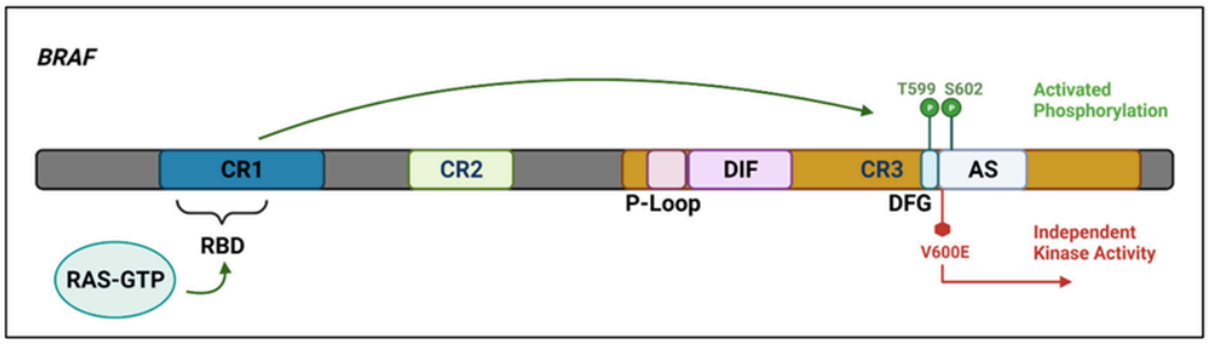


Figure 7: *Schematic representation of the BRAF protein, showing key regulatory domains (CR1–CR3),*

the RAS-binding domain (RBD), and the activation segment (AS). The V600E mutation, located in the DFG motif, leads to constitutive kinase activity independent of RAS signaling (from [26]).

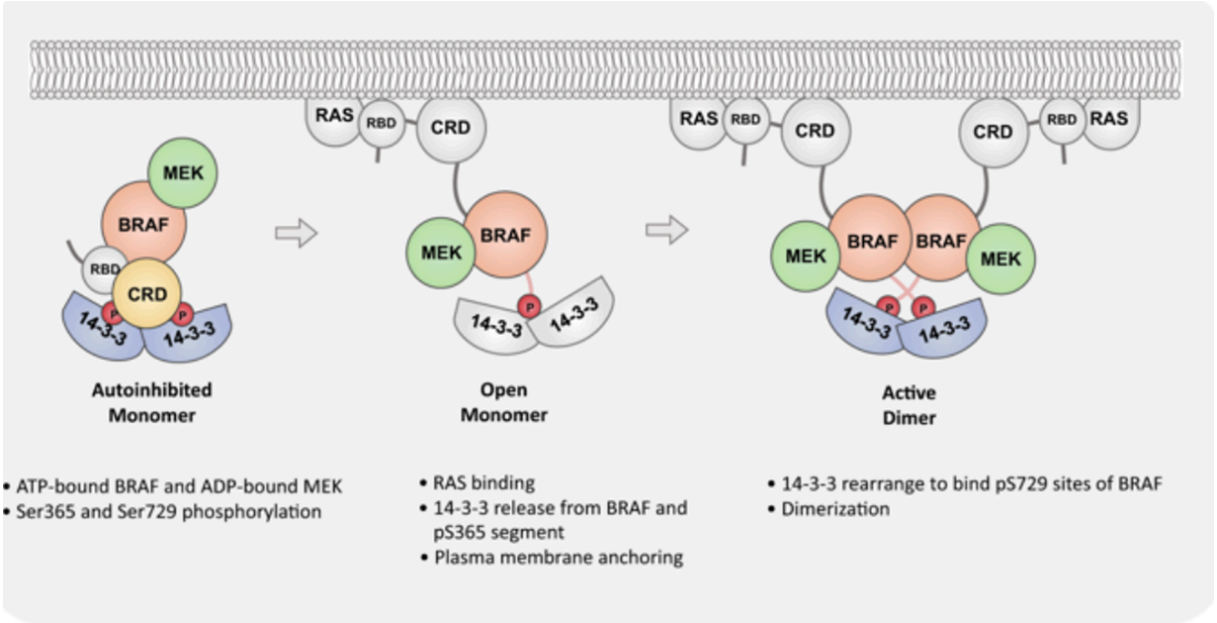
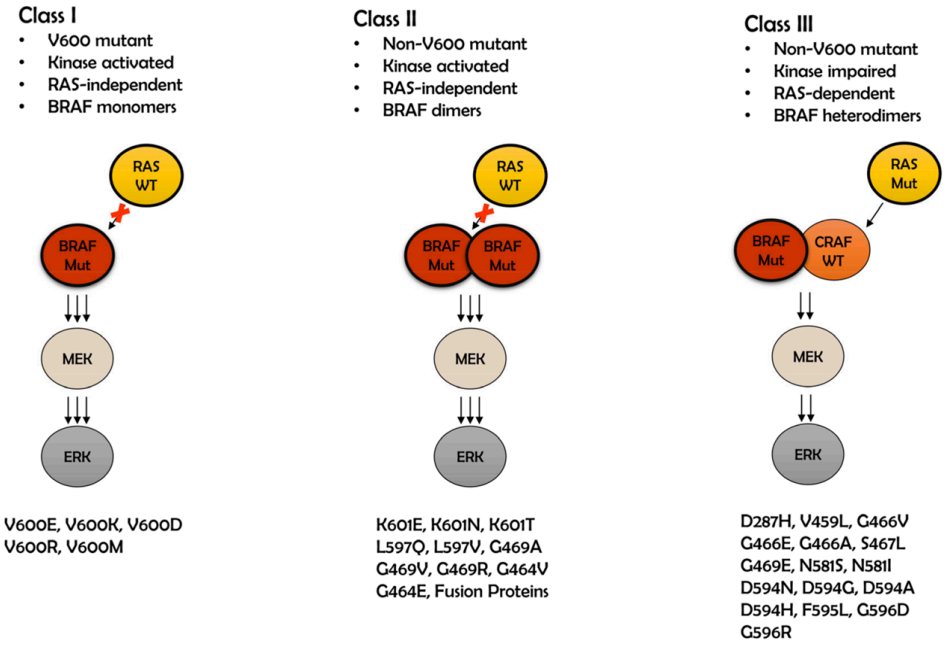


Figure 8: *Activation mechanism of BRAF, the autoinhibited monomer releases the 14-3-3 protein after RAS binding. Dimerization and reorganization of 14-3-3 proteins enable BRAF activation and downstream MEK phosphorylation. (from [24])*

In the literature, Magdalena Smiech & colleagues [22] proposed a classification of different types of mutations of the BRAF protein that lead to over-activation of MEK1 and MEK2 that are commonly seen in many types of cancers. The classification revolves around 3 distinct classes of mutation illustrated in figure 9a: i) Class I mutations are the most common, accounting for over 90% of BRAF mutations in cancer. They involve substitutions at the valine 600 residue, most commonly V600E in the CR3 domain, as shown in figure 9b. This alteration, among others, introduces a negatively charged glutamate side chain that mimics phosphorylation within the activation segment of the kinase. This, in turn, stabilizes the DFG-in active conformation changing BRAF into an active monomer that no longer requires dimerization to phosphorylate MEK. The V600E mutation alone increases BRAF kinase activity by several hundred-fold compared to the wild-type protein, leading to persistent ERK activation and uncontrolled proliferation. ii) Class II mutations are characterized by better dimerization and RAS-independent signaling. Examples

include substitutions like K601E, L597Q, and G469A, which are located either in the activation segment or in the phosphate-binding loop (P-loop). These mutations destabilize the inactive conformation and favor the dimer formation. Class II mutations are maintaining intermediate kinase activity compared to Class I mutations. iii) Class III mutations are variants that paradoxically depend on RAS activity. These mutations, such as D594G, G466V, and N581S, reduce BRAF's intrinsic kinase activity but increase its affinity for RAS-GTP and facilitate activation of CRAF heterodimer formation. In these cases, signaling is maintained or even amplified by CRAF activity in the presence of active RAS.



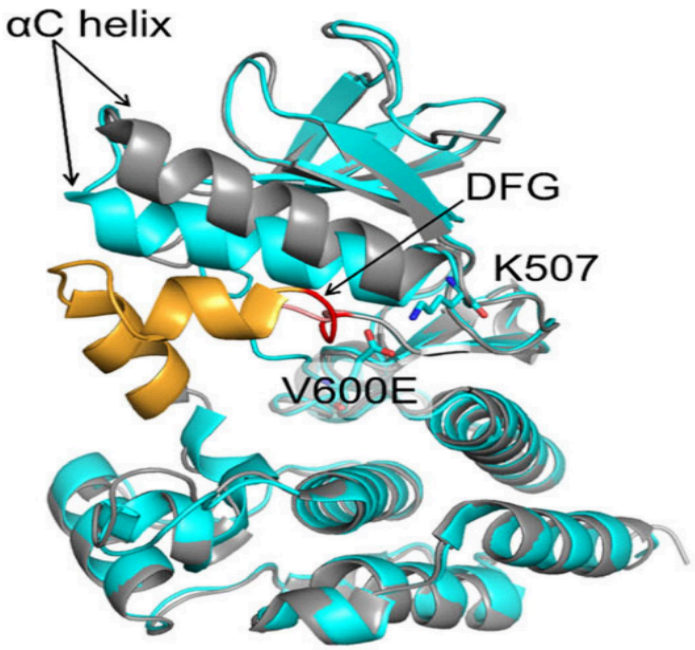


Figure 9: *The top panel illustrates the three major classes of BRAF mutations. Class I includes V600 mutations situated in the CR3 domain. Those are in an RAS-independent activated conformations and function as monomers. Class II mutations are non-V600, kinase-activated, RAS-independent, and signal as dimers. Class III mutations require RAS activation and act as heterodimers with CRAF. Bottom panel represents the structure of BRAF in active form (cyan) after V600E mutation (PDB 4MNF) and closed conformation (gray and orange) (PDB 3TV6), the DFG domain is shown in red (from [22])*

The pharmaceutical recommendations for targeting BRAF have evolved over the past decade, with generations of inhibitors developed to suppress aberrant MAPK signaling [figure 10]. The first clinically approved drugs, vemurafenib and dabrafenib, selectively inhibit the active monomeric form of BRAF V600E and show strong benefits in melanoma. However, resistance often develops within months of treatment. Resistance mechanisms include upregulation of RAS, alternative splicing of BRAF producing variants with enhanced dimerization capacity, and activation of parallel survival pathways like PI3K-AKT. To address these challenges, combination therapies were developed that pair BRAF inhibitors with MEK inhibitors (e.g trametinib or cobimetinib) to block the pathway at multiple levels. This combination strategy tends to delay resistance development. Further strategies tried inhibitors able to target RAF dimers directly and overcoming dimer-dependent resistance. More

recently, ERK and RAF inhibitor combinations were designed to suppress the feedback reactivation of the pathway.

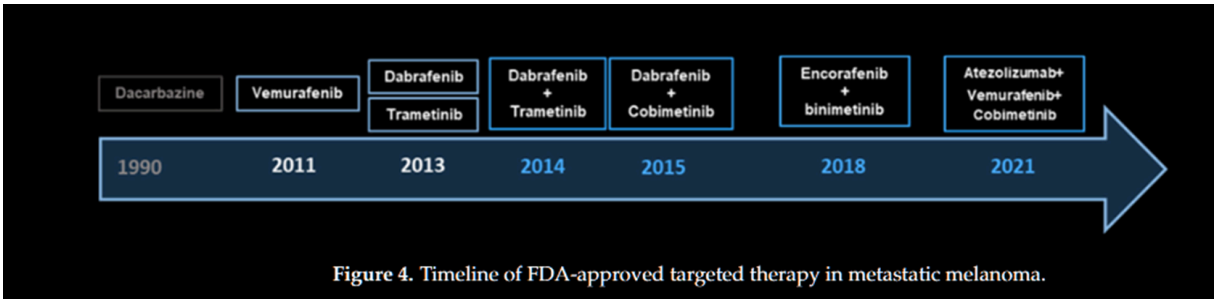


Figure 10: *Timeline showing the evolution of drug therapies for targeting BRAF V600E mutation in metastatic melanoma (from [21])*

1.6 State of the Art

The development of huge clinical databases like GDSC or CCLE, referencing many aspects of cancer cell lines biology, has had a huge impact on approaches trying to predict drug efficacy for specific cell lines. Those databases represent a new paradigm where highly accurate and automated experiments on thousands of fully characterised cell lines can be experimentally tested on a huge arsenal of drug compounds, in turn leading to better and more effective treatments tailored to the patient's unique cancer features. Those valuable datasets encourage the research field to develop prediction algorithms aiming to predict in advance the sensitivity of cell lines for different therapeutic compounds. The literature shows many interesting approaches that have been implemented by researchers in the field using the data available in the GDSC database. [Table 1] shows a non-exhaustive summary of different approaches using specific features and prediction algorithms. These methods propose a wide range of strategies, from classical statistical tests to machine learning classifiers and optimization functions. Some approaches focus on

identifying gene expression markers associated with sensitivity or resistance [29,30], while others use predictive classifier models trained on mutational or multi-gene signatures to classify cell lines into resistant or sensitive categories [28,29,31]. Many of these tools use genomic features such as point mutations, copy number variations, or expression data and use drug response metrics. In [28] paper's approach, authors use typical machine learning models (SVM, KNN, …) to classify cell lines and achieve 90% accuracy. Those are strong results but they lack reproducibility in experimental conditions. Other approaches have tried to predict cancer cell lines drug-sensitivity values (IC50) using regression algorithms [32], carefully comparing different features and prediction model architectures for finding the best strategies. In papers [29,30], authors achieve to discover new genetic and expression markers for BRAF V600E resistance. However, current models still face important limitations. A common issue for machine learning models is overfitting the training data, where models fail to generalize the information on unseen data and predictions correspond too closely to the training dataset. In such a case, the model fails to generalize to additional data or predict future observations reliably. Another limitation is that most models do not take into account structural information about protein–drug interactions. This is especially important for personalized oncology where features like mutations or expression only give a partial biological explanation for sensitivity or resistance mechanisms. Structural prediction models might close the gap between statistical correlation and biophysical consequences of protein dynamics. In contrast, this paper tries to leverage the power of mutational or statistical approaches by exploring new structural features produced by state-of-the art deep learning structure prediction methods like Alpha Fold 3.

ARTICLE REF	YEAR	TYPE OF PREDICTOR	TARGET	FEATURES	ALGORITHM	RESULTS
A novel heterogeneous network-based method for drug response prediction in cancer cell lines	2018	Synthetic approach information flow-based algorithm	IC50 (2 classes: resist/ sensitive)	protein-protein interractions cell line genomic profile drug chimical structure (1D, 2D not 3D) drug-target interraction	heterogeneous network-based method (HNMDRP)	Comparaison des AUC avec d'autres algorithmes
A tool for discovering drug sensitivity and gene expression associations in cancer cells	2017	Statistical test	gene expression markers associated with drug sensitivity/resistance	Gene expression data (RNA-seq from CCLE and GDSC). Drug sensitivity metrics (IC_{50})	Spearman's correlation analysis	identify HGF, MET, and VEGF-A expression correlations with resistance to a BRAF(V600E) inhibitor
Precision and recall oncology: combining multiple gene mutations for improved identification of drug-sensitive tumours	2017	comparison between multi gene marker, machine learning algorithm and signlegene markers	IC50 (2 classes: resist/ sensitive)	signle gene mutations or multi gene markers	ML Classification algorithm	Multi-gene markers outperform single-gene markers
Comparative analysis of regression algorithms for drug response prediction using GDSC dataset	2025	Compare 13 regression algorithms and 3 feature selection methods on GDSC	prediction of drug response, (IC50)	Gene expression, Mutations, Copy Number Variations & Drug group	13 different scikit learn algorithms	SVM is best and CNV has little effect in sensitivity
Prediction of cancer cell sensitivity to natural products based on genomic and chemical properties	2015	Machine learning (different models: knn, SVM, randm forest, rotation forest, Descision tree)	IC50 (2 classes: resist/ sensitive)	Genomic features (GDSC) Chimical 1D and 2D descriptors (SMILES)	4 different classification ML algorithms	Accuracy between 60 and 90% in general

Table 1: *List of recent computational approaches for predicting cancer drug response using the GDSC database. The table summarizes various methods by publication year, prediction type, target data, key features used, algorithmic strategies, and reported outcomes.*

Chapter 2 – Objectives

The main objective of this thesis is to assess if the integration of protein structural information into drug response prediction pipelines can help in the identification of sensitive or resistant cancer cell lines to specific compounds.

More specifically, objectives can be described as following:

1. Manipulate genomic, transcriptomic, and drug sensitivity data from the GDSC database in a scientifically relevant setup.
2. Use deep learning tools such as Alpha Fold 3 and Boltz 2 to predict 3D structure of protein with high precision.
3. Assess how mutations in BRAF proteins can modify its 3D conformation, changing its sensitivity to specific anticancer drugs.
4. Evaluate drug–protein binding affinity using structure-based predictors like Boltz 2.
5. Integration of multi-omic features in the drug sensitivity prediction model to enhance drug sensitivity predictions.

All those objectives aim to explore and evaluate the contribution of structural bioinformatics in predictive oncology models.

Chapter 3 – Methods

Alpha fold models

In 2018 Google DeepMind in collaboration with the biotechnology company Isomorphic labs started a revolution in the field of structural biology by releasing a new protein folding prediction algorithm. This novel technique was using the capabilities of deep neural networks to try to solve the complex problem of predicting the three-dimensional structure of a folded amino acid sequence of any length. The mechanisms by which an arbitrary long chain of peptides could form secondary and tertiary structures and fold itself into a thermodynamically favorable conformation was still poorly understood. In 1994, an international contest was created to let research teams compete in trying to predict protein structures blindly. This contest was called "Critical Assessment of Protein Structure" or CASP. In 2018, the Google model 'Alpha Fold' won the '*ab initio*' section of the CASP challenge. Contrary to other sections, the '*ab initio*' section demands to predict new protein structures using no template or prior knowledge to make the final prediction. Results in this section gained progress over a decade and led to the introduction of Alpha Fold 2 techniques in the following round of CASP (2022), which achieved very high accuracy in protein structure prediction. With the success of a number of methods that build on the ideas and techniques of Alpha Fold (RosettaFold, Alpha Fold 2-multimer, etc…), lead to the question of whether it is possible to accurately predict the structure of complexes containing a much wider range of biomolecules. A few years later, DeepMind released their new model Alpha Fold 3 (AF3) proposing new interesting features like

ligand or complex binding prediction and even interactions with DNA or RNA or other small molecules. Looking closer at the difference in model architecture between the 2 latest versions of Alpha Fold [figure 11] we see that the general procedure for generating structure predictions has been amended, adding key modifications to enable multi-molecular prediction with much higher accuracy compared to specialized tools for protein-ligand interactions, like state-of-the-art docking tools or better predictions for antibody antigen predictions than alpha fold multimer v2.3 [33]. Key changes between alpha fold 2 and 3 include:

1. Diffusion model replaces structure module to allow more flexibility and stably during the generation process
2. Alpha Fold 3 does not rely as much on Multiple sequence alignment (MSA) allowing it to predict other molecules or engineered proteins.
3. Table 2 compares the performance metrics used in Alpha Fold 2 and 3.

One of the major issues with generative models is that they can hallucinate, producing plausible but incorrect structural regions. This is common in unstructured segments. Additionally, these predictions can lead to predicted structures displaying elongated, “spaghetti-like” conformations when uncertainty is high and model confidence is low. In terms of stereochemistry, AF3 predictions have two main limitations: the model sometimes produces chirality errors, and it can generate atomic clashes, particularly in large protein complexes with over 2,000 residues. Another limitation is that AF3 predicts static structures, similar to PDB snapshots, rather than capturing the dynamic range of conformations found in solution.

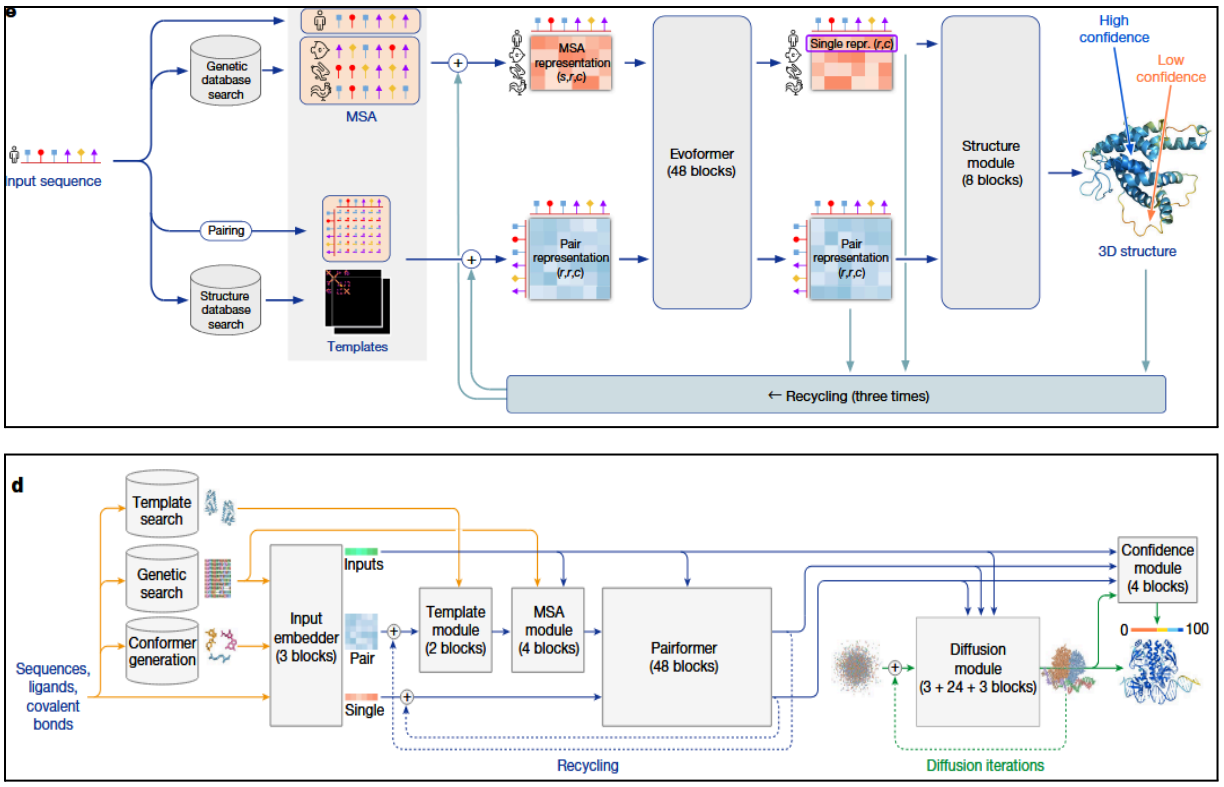


Figure 11: *(a) Architecture of Alpha Fold 2: uses genetic and structural databases to build multiple sequence alignments (MSA), processes these with the Evoformer and Structure Module.*
(b) Architecture of Alpha Fold 3 removing the alignment module. Unified architecture with a Pairformer and a Diffusion Module instead (from [34,35])

Metric Name	Use	Availability
PAE (Predicted alignment error)	2D matrix showing positional error for each residue i when it is aligned on residue j.	AF2 & AF3
pLDDT (per residue Local Distance Score)	Score between 0 and 100 estimate the position accuracy of the backbone at each residue	AF2 & AF3
pTM	Predicted measure of structure similarity when compared to ground truth	AF2 & AF3
ipTM	Predicted score indicating how confident the model is for the interface between multiple chains	AF2-m & AF3
Model confidence	Internal ranking score selecting the best model from multiple runs	AF3

Table 2: *Description and Comparison of Alpha Fold 2 (AF2), Alpha Fold 2-multimer (AF2-m) and Alpha Fold 3 (AF3) confidence metrics*

Boltz 2

Many tools have been developed in recent years to predict protein structures and how complexes fold themselves under different chemical conditions. Reliable predictions would greatly accelerate costly drug discovery processes that require screening thousands of compounds. Advances like Alpha Fold 3 have improved structural modelling but still cannot estimate binding strength between candidate drugs and target proteins. To address this, Passaro and colleagues introduced Boltz 2 which is a deep learning model that combines structural prediction and quantitative binding affinity estimation in a single framework [36]. Boltz 2 was inspired by Alpha Fold 3 and has been designed to help with drug discovery. Boltz 2's approach leverages deep learning and energy-based methods, claiming accuracy comparable to state-of-the-art free-energy perturbation (FEP) simulations in predicting protein-ligand binding affinity. Its affinity prediction module is trained on a curated dataset of biochemical data, including dissociation constants (K_d) and IC50 values collected from public assays. The model learns to classify binders versus non-binders and predict IC50 with good results compared to experimental measurements. [36]

Structural visualizations and analysis with PyMol

In this section we describe how we have visualized different conformational structures of BRAF proteins that were visualized with PyMol version 1.74, a molecular visualization system created by Warren Lyford DeLano [37]. We started by extracting BRAF structures in both the active and inactive conformations from the Protein Data Bank, and in detail, we used the following structure:

- **PDB 3OG7**, corresponding to a crystal structure of BRAF V600E bound to the selective inhibitor PLX4032 (vemurafenib analog)

- **PDB 3C4C**, representing a BRAF Wildtype kinase domain in an inactive conformation complexed with a different inhibitor.

- **PDB 4XV2,** corresponding to a crystal structure of BRAF V600E bound to the selective inhibitor dabrafenib

A script was developed to load multiple protein structures and automatically apply distinct color themes for easy differentiation. Once loaded, all molecules are aligned using PyMOL's `align` function, ensuring consistent orientation across the 3D space. The script supports both experimentally resolved structures from the PDB and predicted models from Alpha Fold or Boltz. To highlight key features, residue 600 is automatically colored in red, and interface residues were manually annotated in orange.

BeAtMuSiC

BeAtMuSiC [40] is a computational tool developed by our laboratory to assess the impact of amino acid substitutions on protein-protein interaction. The primary goal of this program is to estimate the change in binding free energy (ΔΔG) following a mutation.

$$\Delta\Delta G = \Delta Gmutant - \Delta Gwt$$

where ΔG represents the binding free energy of the variant (mutant, wt) to a specific drug.

It is based on a series of statistical potentials, specifically mean force potentials extracted from resolved protein 3D structures, which are appropriately combined. The tool is available through a web interface that allows users to submit PDB structures of the target protein they wish to mutate.

PremPLI

While BeatMuSiC computes protein–protein interactions, PremPLI [41] evaluates the impact of mutations on protein–ligand binding affinity. It integrates evolutionary information along with structural data on the protein-ligand complex to assess how mutations can affect the binding between a protein and its ligand. PremPLI is also accessible via a web server, where users can upload their complex of interest and perform mutagenesis predictions.

Alpha Fold 3 clustering procedure

We generated Alpha Fold 3 predictions using the open-source version of the program available at https://github.com/google-deepmind/alphafold3. With help from some colleagues, we ran it on our lab's server, which has an AMD processor (64 cores) and four NVIDIA A40 Ampere GPUs, giving unrestricted access to the program via the command line. To run the program, we provided the input sequences of the molecules in a JSON file.

We used Alpha Fold 3 to analyze BRAF, generating different protein conformations: 10 for the wild-type BRAF and 10 for the BRAF V600E mutation, with seed values ranging from 1 to 10. We then used PyMOL to align all 10 conformations (as shown in subsection 4.3). After selecting the best model prediction as the cluster centroid, based on Alpha Fold's confidence metrics, we calculated the per-residue RMSD for each model relative to the centroid residues. Finally, we average these RMSD values across the 10 models to get the average cluster RMSD for each residue.

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[(x_i^A - x_i^B)^2 + (y_i^A - y_i^B)^2 + (z_i^A - z_i^B)^2\right]}$$

where N is the number of atom pairs included in the calculation; x, y, and z are the Cartesian coordinates of each atom in structures A and B.

Finally, by comparing the average RMSD profiles of the wild-type and mutant clusters residue by residue (taking the absolute difference at each position), we obtained the

comparative profile shown in figure 20. This analysis has been done to gain insights into how V600E mutations impact the BRAF structure.

Correlation of Affinity Prediction with Experimental log(IC50) values

Unique Single Nucleotide Variants in *BRAF* observed in the combined GDSC databases [38] were extracted. Indels, frameshifts, duplication deletion and non-coding mutations were filtered out. All the Simplified Molecular Input Line Entry System (SMILES) codes of the available drugs in combined GDSC databases were obtained from the PubChem REST API. Compounds with missing SMILES were discarded. The affinity prediction of all the mutation / SMILES combinations were computed using Boltz 2 (2.0.0).

The spearman correlation between these affinities predictions and to the log(IC50) experimentally obtained from the GDSC database was computed. Spearman correlation was preferred over Pearson due to the non-normalized distribution of the data and to reduce outliers contribution to the score.

Boltz 2 drug discovery pipeline & correlations

For the drug discovery pipeline, we utilized Boltz 2 to predict the affinity of many cancer drugs to our target gene BRAF in a similar way as in the previous section: we started by writing a script to crawl the SMILES codes for all drugs present in the GDSC database by aggregating GDSC 1 and GDSC 2 in a combined dataset and extracting all unique drug names. Input templates were generated for each unique compound to assess the binding strength using Boltz 2. We ran a script to launch each drug template on the lab's server using Boltz 2. In the meantime we extracted the log(IC50) values from GDSC for corresponding cell lines by averaging over cell lines sensitivity values. We generated a csv file to store the log(IC50) values from GDSC experimental data, from Boltz 2 predictions for wildtype and for V600E mutant.

This file was then used to compute spearman correlations between predicted and target sensitivity values.

Expression data integration

Here we basically collect for each cell line the expression data for each gene. This transcriptomic data was retrieved from the cell model passport web portal [39] using the '`all RNA-Seq processed data (1.02 GB)`' dataset. The integration of expression data was done via an R script, and the differential analysis was performed using the *DESeq2* [42] library with default parameters. To reduce overall noise, we filtered out all genes with a count below 10. Top 10 genes were selected based on |*log(Foldchange)*| between wild type and V600E conditions keeping only significative (*p*-value adjusted < 0.05) differences. For each drug, a final expression score was computed by taking the median expression of these selected genes across all cell lines tested with that drug. This expression feature was then used as an additional variable in the linear model in combination with the Boltz 2 structural prediction scores.

Chapter 4 – Results

4.1 GDSC dataset description

Here, I will describe the GDSC database to help us better understand its organisation by showing and exploring some important features it provides. The database is composed mainly of three different sections containing information on the compounds and cancer cell lines, along with one major output: the logarithm of the IC50 value, which indicates drug sensitivity levels of a particular cell line to a specific drug. The 3 sections one can browse and search are the ‘Features’ , ‘Compounds’ and ‘Cell lines’. The ‘Features’ section contains genetic signatures that have been predicted by the database authors using a deep learning model to play a role in cancer drug sensitivity or resistance changes. Regarding the ‘compound’, the GDSC database references over 624 different chemotherapeutic compounds and tests the effects of the vast majority of them on all the cell lines using high-throughput techniques. [Figure 13a] shows that the vast majority of drugs have been tested on at least 80% of the cell lines. Only five compounds (THZ-1-87, THZ-2-98-01, XMD11-50, HG-6-71-01, Torin2) have been screened against less than 25% of the cell lines, for reasons that the authors of GDSC do not explicitly specify in their paper [19]. Drugs that will be analyzed in this paper, like dabrafenib, Quizartinib or Methotrexate were respectively in 99.5%, 93.5% and 99.3% of all 978 cell lines. The database is separated into two distinct datasets called *GDSC1* and *GDSC2*, where *GDSC2* represents more recent data (after 2015). The distribution of drugs and cell lines between GDSC1 and GDSC2 is shown in figure 14. For this thesis, I decided to systematically aggregate the combined information of GDSC1 and GDSC2 to incorporate the entire database in all our results without specific filtering. When multiple sensitivity values were present for the same drug/cell line combination, I averaged their IC50 values in the final analysis.

Each section has been further annotated with a few important informative factors. For example the ‘Cell line’ section contains identifiers to relate them to databases like cell passport identifier or the COSMIC ID. We also have information about the tissue and tissue subtype from which the cell line originates. Another important attribute is the TCGA classification for each cell line [figure 13b]. The TCGA classification is a cancer classification standard developed by *The Cancer Genome Atlas* (TCGA) project, which comprehensively characterized the molecular features of thousands of tumors across many cancer types. It groups cancers based on genomic, transcriptomic, and epigenetic profiles rather than only their tissue of origin [6].

[Figure 13c] shows the proportion of each TCGA class in the GDSC 1 and 2 datasets. The compounds also have a few important attributes we can refer to, mainly the *'target'* and *'target_pathway'* attributes, that represent respectively the molecular protein target of the anti-cancer drug and the molecular pathway this target is part of. For example, dabrafenib, one of the drug we have analyzed more in detail, has BRAF as its main target and the related pathway is *ERK MAPK signaling*. [Figure 13d] shows how those pathways are distributed in the GDSC database.

Additionally, the database website incorporates many additional features like genetic variations, copy number aberration, epigenomics and expression data of cancer cell lines that can be used to better understand their role in drug response mechanisms. For this thesis, we limited the exploration to two of the most common features in the literature: genetic variations and expression profiles.

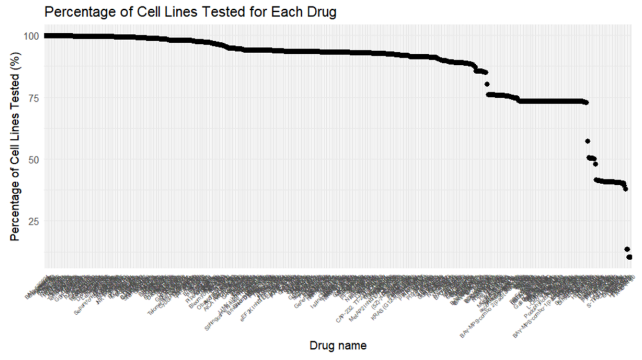
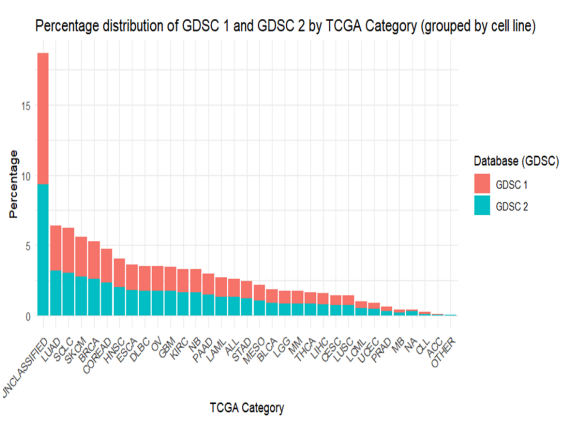


Table of TCGA Cancer Codes

TCGA code	Cancer type	Number*
ACC	Adrenocortical carcinoma	90
BLCA	Bladder urothelial carcinoma	396
BRCA	Breast invasive carcinoma	988
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	194
CHOL	Cholangiocarcinoma (bile duct)	35
COAD	Colon adenocarcinoma	367
DLBC	Lymphoid neoplasm diffuse large B–cell lymphoma	48
ESCA	Esophageal carcinoma	185
GBM	Glioblastoma multiforme	290
HNSC	Head and neck squamous cell carcinoma	513
KICH	Kidney chromophobe	66
KIRC	Kidney renal clear cell carcinoma	451
KIRP	Kidney renal papillary cell carcinoma	282
LAML	Acute myeloid leukemia	197
LGG	Brain lower grade glioma	533
LIHC	Liver hepatocellular carcinoma	373
LUAD	Lung adenocarcinoma	546
LUSC	Lung squamous cell carcinoma	178
OV	Ovarian serous cystadenocarcinoma	469



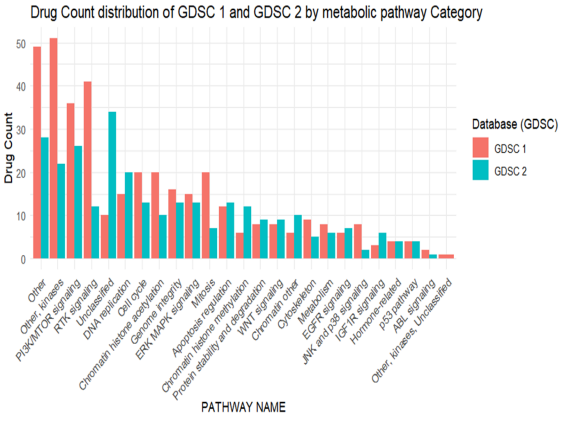


Figure 13: *(a) Top left shows the percentage of cell lines covered in GDSC for each drug present in the dataset. (b) Top right shows a table indexing all TCGA codes and their signification. (c) Bottom left percentage of cell lines GDSC 1 (orange) and GDSC 2 (blue) classified by TCGA code. (d) Bottom*

right counting all drugs in GDSC 1 (orange) and GDSC 2 (blue) depending on the pathway of their target protein.

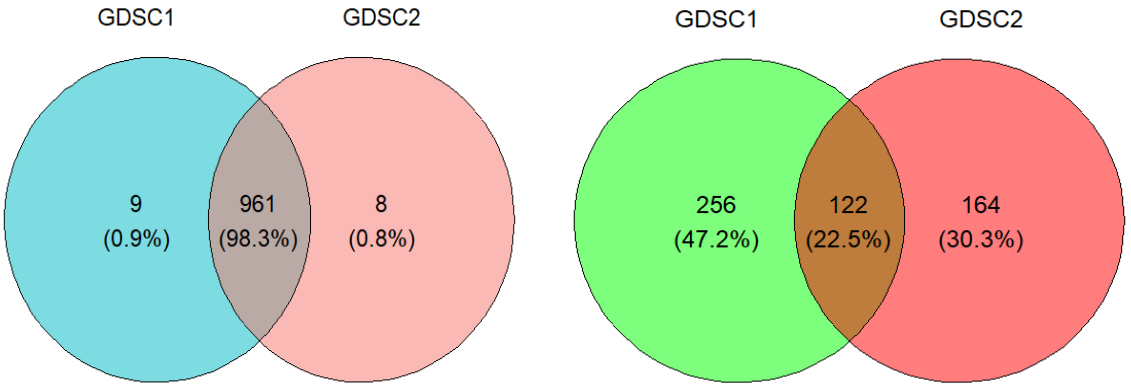


Figure 14: *(Left) Venn diagram showing the distribution of cell lines across both GDSC1 and GDSC2. Showing that 98.3% of cell lines are common between both databases. (Right) diagram shows drugs distribution between GDSC 1 and 2 with only 22.5% drugs common to both datasets.*

4.2 Genetic variations

Mutations are one of the most important features when looking at drug resistance because they frequently occur and have wide and sometimes complex effects on drug sensitivity or cancer progression. Mutations in drug target genes can profoundly affect drug sensitivity by modifying protein conformation and function. Comprehensive databases characterizing mutations and their effect are useful tools for researchers to identify mutations and their effect in different cancers. These databases also become important for developing predictors of which drugs would have better anti-tumoral effects in a large panel of potential drug candidates, thereby helping to identify possible molecular leads. We have shown in the previous chapter some machine learning models that take mutational status into account to predict drug efficacy by different methods. In this work, we aim to investigate this mutational relationship in a more biophysical way by focusing our predictions on the mutation effects on the drug-protein complex, rather than analyzing the whole cell. To illustrate this idea, we focused on the protein BRAF from the RAF protein family, a key protein in the MEK/ERK pathway regulation. Many kinase inhibitors are known to specifically

target the BRAF V600E while being less sensitive to wild type versions of the BRAF protein. Databases such as GDSC report sensitivity parameters that corroborate the research. In figure 12, we plotted the mutational status of each cell line in the database in relation to its sensitivity to dabrafenib, a common type I kinase inhibitor used to treat BRAF V600E patients, which binds to the inactive conformation of BRAF and stabilizes it. As expected, we observe a significant difference in drug sensitivity between the cell lines with BRAF carrying V600E, BRAF wild type, and BRAF carrying other mutations. This clear separation in IC50 distributions demonstrates the strong impact of V600E on drug sensitivity to dabrafenib. The same analysis on another drug, Methotrexate, a DNA synthesis inhibitor, shows no statistical difference in sensitivity across different mutational status groups. Looking at overall statistics we see that BRAF V600E mutation is present in 14.7 % of all cell lines and appears principally in Skin Cutaneous Melanoma (SKCM) and Thyroid Carcinoma (THCA) which is coherent with the literature. Those results are an example that the GDSC database is a great resource to identify interesting mutational features that can be used to predict drug sensitivity. For example, in table 3, we show the 10 drugs that are most sensitive to BRAF V600E in the GDSC database. The sensitivity is measured by the difference in log(IC50) values between cell lines with and without the mutation. Note that not all of them target BRAF specifically.

In the next sections, we incorporate mutational information at the level of the target-drug molecular interaction, exploring how the structural changes in the protein structure induced by mutations can explain how BRAF mutation has such a drastic impact on the sensitivity to specific drugs like dabrafenib.

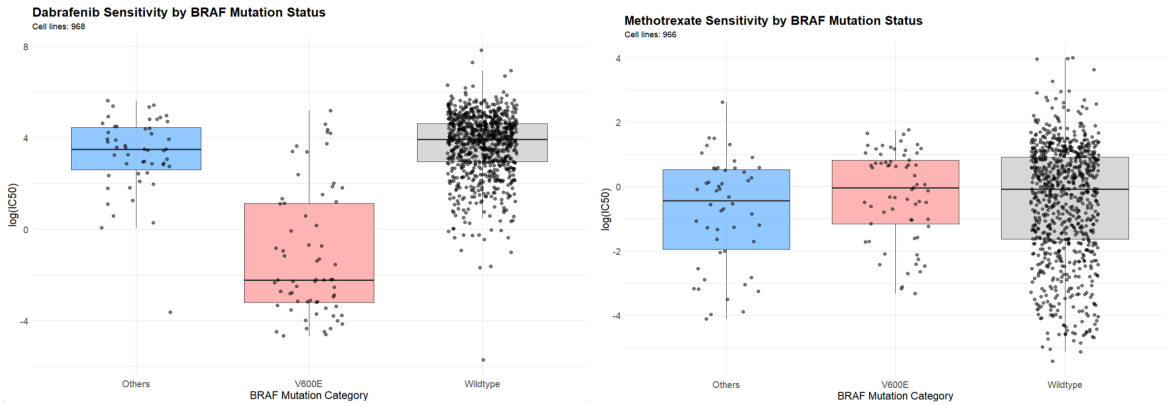


Figure 12: *Comparison of drug sensitivity (log_{10}[IC50]) across BRAF mutation categories for dabrafenib (left) and methotrexate (right). Dabrafenib shows significantly increased sensitivity (lower IC50) in V600E-mutant cell lines compared to wild-type and other BRAF mutations.*

Drug	Drug Target	Effect size	P-value
Dabrafenib	BRAF	-2.87	3.07e-47
PLX-4720	BRAF	-2.19	1.8e-31
PLX-4720	BRAF	-1.95	1.07e-24
SB590885	BRAF	-1.92	1.41e-21
LIMK1 inhibitor BMS4	LIMK1	-1.09	3.71e-13
RAF_9304	ARAF, BRAF, CRAF	-1.35	5.13e-11
(5Z)-7-Oxozeaenol	TAK1	-1.44	3.92e-10
VX-11e	ERK2	-1.02	6.58e-10
Selumetinib	MEK1, MEK2	-1.39	2.17e-07
CI-1040	MEK1, MEK2	-1.2	8.43e-07

Table 3: *Top 10 drugs showing differential sensitivity in BRAF V600E-mutant cell lines. Effect sizes indicate stronger drug efficacy (lower IC50) in V600E mutation versus wild-type (from https://www.cancerrxgene.org/feature/BRAF_mut/50/volcano?screening_set=GDSC2).*

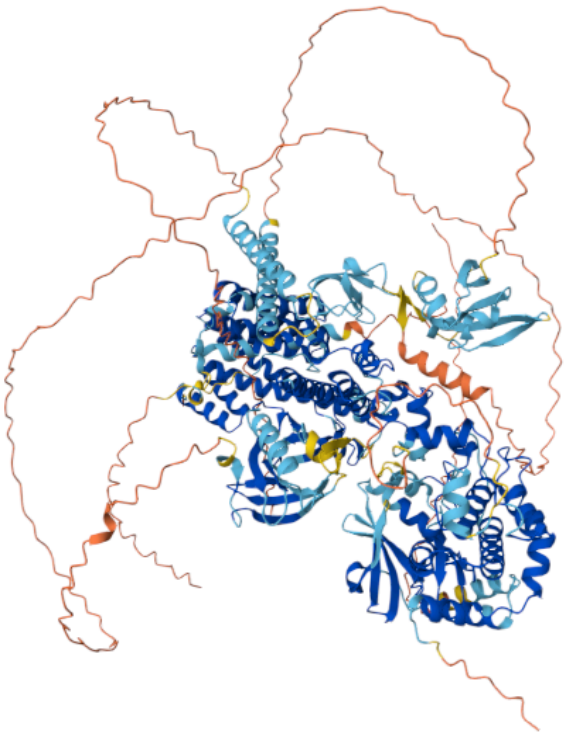
4.3 Structural analysis

4.3.1 Alpha Fold 3 predictions

In this report, we aim to evaluate if specialized deep learning tools like Alpha Fold 3 or Boltz 2 might be used to predict cell sensitivity to cancer drugs reliably. The assumption is that finding a quantitative proxy for the binding affinity between the target protein and drug compound might be achieved by using the Alpha Fold 3 model's advanced structural prediction capabilities. To evaluate this idea, we selected the BRAF gene due to the characteristics described in Chapter 1, which outlines the ERK/MAPK signaling pathway and the role of BRAF as a key regulator in cancer survival and proliferation.

First we started by identifying all the molecular actors that interacted with the BRAF protein. The idea is to identify which ones would be affected by the conformational change induced by the V600E point mutation. Once identified, we started by running a prediction for BRAF protein structure using Alpha Fold 3 webserver [43]. The results in figure 15 shows the predicted BRAF structure in an auto-inhibited state bound to MEK and 14-3-3. Even when starting from a pdb structure, in this case the 6Q0J pdb entry, alpha fold predicts some parts of the protein in a nice, structured way but has huge unstructured filaments predicted with low confidence often referred to as spaghetti-like structures. Those are the signs of intrinsically disordered regions and might be a clue that the Alpha Fold 3 prediction algorithm is not a reliable tool to predict multi-molecular complex interactions between BRAF, RAS, CRAF, MEK.
Second we studied the BRAF protein and the drug dabrafenib using Alpha Fold 3. The removal of all the other molecular interactions like RAS, MEK or 14-3-3 permitted us to isolate only the binding with the drug but it omitted a lot of molecular context that might be relevant in-vivo. To further enhance the predictions of the model, we found that isolating the RAS binding domain of the BRAF protein with the target drug dabrafenib, resulted in much more accurate and consistent predictions compared to PDB structures and resulted in better confidence scores [figure 16].

a)



b)

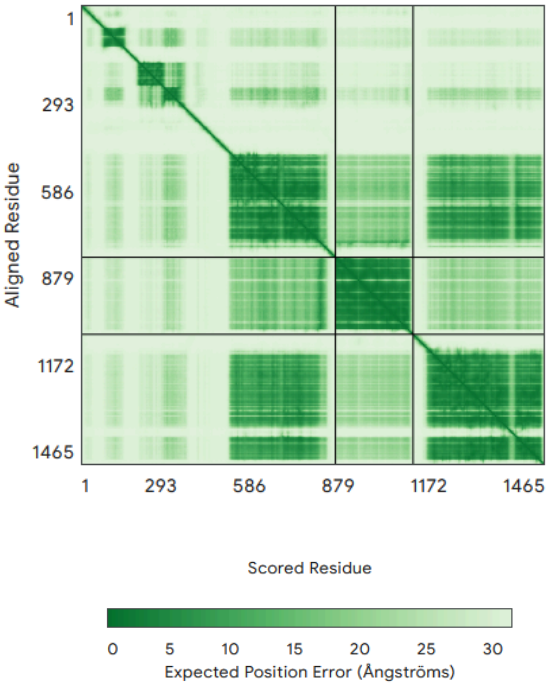
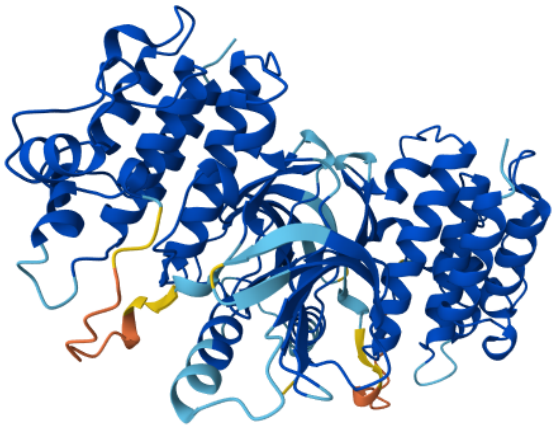


Figure 15: *Alpha Fold 3 server structure prediction for the full sequence BRAF protein in an auto-inhibited state bound to 14-3-3 and MEK protein. a) The coloring of the complex indicates the model's confidence at each residue. Blue indicates high confidence orange indicates lower confidence. b) The green box indicates the expected position error in Ångström after alignment.*

a)



b)

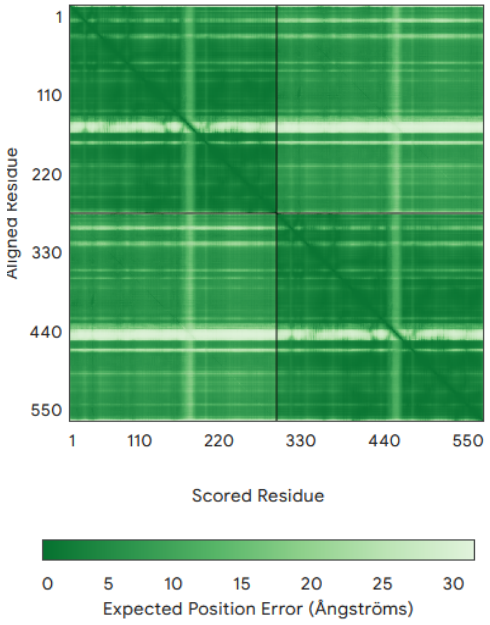


Figure 16: *Alpha Fold 3 server structure prediction for the CR3 domain of two BRAF proteins. The coloring of the molecule indicates the model's confidence at each residue (pLDDT). The green box indicates the expected position error in Ångström after alignment.*

We then analyze the generated structure in PyMOL to better characterize and compare the predicted conformations of the wild-type and mutant BRAF proteins with experimentally resolved crystal structures. This comparative approach was done to assess whether the in-silico models captured conformational features consistent with known active and inactive kinase states and to identify potential structural explanations for the differences in drug sensitivity observed in cell lines. The visualization workflow first involved loading Alpha Fold-predicted structures of BRAF wild-type and BRAF V600E into PyMOL and aligning them with relevant reference structures obtained from the Protein Data Bank (PDB). Three reference BRAF structures were selected for this comparison (3OG7, 3C4C, 4XV2, see methods). On figure 16, visual inspection shows that both predicted models are aligned closely with the reference structure in the N-terminal and C-terminal lobes of the kinase domain. Small differences are apparent in the region corresponding to the activation segment

and the αC-helix. The PDB 3OG7, corresponding to a crystal structure of BRAF V600E, shows a slight displacement of the activation loop compared to the wild-type model, resembling the active state of the kinase. This observation is consistent with the known functional effect of the V600E substitution, which disrupts the regulatory interaction that normally stabilizes the inactive conformation. To further illustrate these structural relationships figure 18 shows a representation in which the predicted BRAF–MEK interaction interface is illustrated. This representation demonstrates that although the V600E substitution does not directly participate in the binding interface with MEK1, it is positioned adjacent to regions critical for stabilizing the active conformation of BRAF, reinforcing the hypothesis that the mutation might have allosteric effects. The visualization also shows the relative orientation of the αC-helix and activation segment, both of which are important determinants of kinase activity and inhibitor binding.

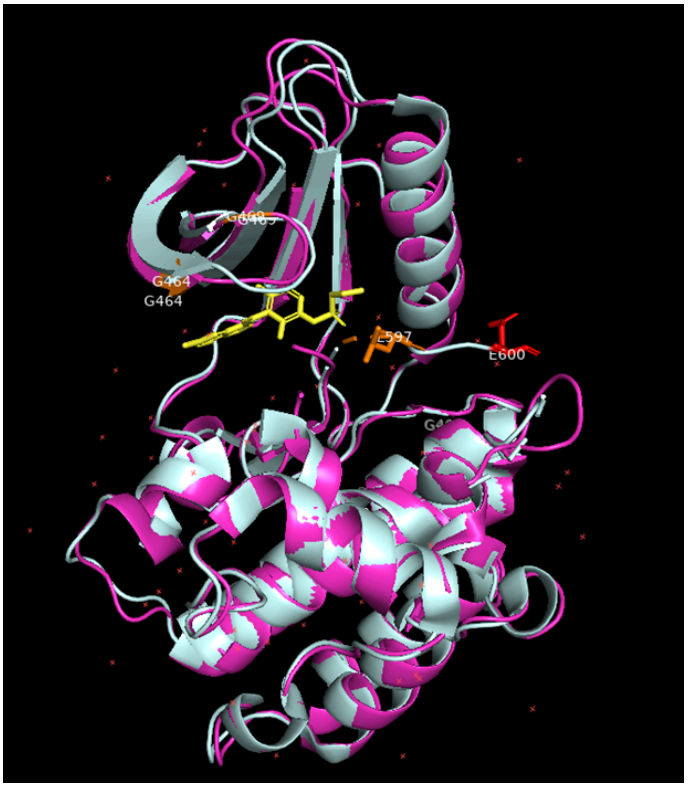


Figure 17: *shows an overlay of BRAF V600E (colored white - PDB 3OG7) and the BRAF wildtype model (colored purple - PDB 3C4C).*

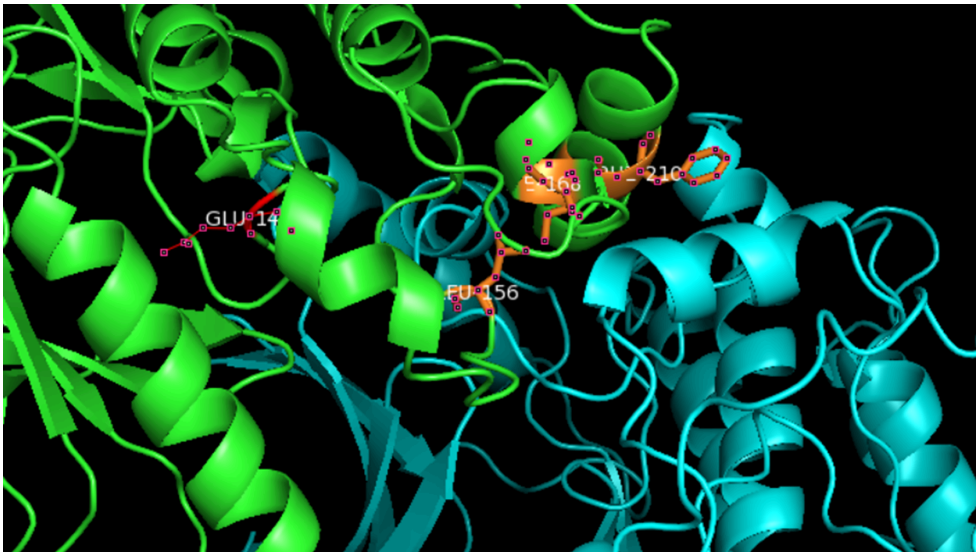


Figure 18: *Alpha Fold 3 prediction for BRAF MEK bound to dabrafenib, BRAF is in green MEK in cyan, interface residues are in orange, the red residue represents the V600E mutation.*

4.3.2 Clustering (RMSD- pLDDT)

In this section we describe how using clustering methods on predicted structures of AF3 lead to discovering important clues regarding the objectives defined in the previous sections. Deep learning models such as AF3 use many randomized events in their architecture, for example the starting parameters of the model or the noise pattern added in the diffusion steps. This means that running multiple predictions using the same input data can lead to a different output. The main technique to avoid incorporating unreproducible randomness into the predictions is to use different seed values. We decided to run each prediction 10 times using 10 different seeds [1 to 10]. Those sets of predictions form an ensemble E of predictions using the same input data with slightly different outputs due to AF3's algorithmic randomness as shown in [figure 19]. The initial conditions we used for this test were the following: i) first predictions includes only BRAF wild type and dabrafenib molecules, where the CR3 binding domain has been isolated to limit the overall complexity of prediction. ii) The second category is the same setup but BRAF has been mutated to the V600E

variant. [Figure 19] illustrates all those different clusters and how much variability there is between them. In fact, apart from long structured strings and loops the different seed predictions are identical for each setup. There is also one interesting phenomenon that appears when you look specifically at the position of residue 600 marked in red on the figure, for each conformation ensemble E, the residue 600 is usually positioned in a stable way that can be characterized as active or inactive depending on if the residue is a Valine or Glutamic acid. The wildtype structures, with the residue 600 valine, have the nonpolar side chain forcing the activation loop in an inactive conformation. This is also the case when predicting with AF3 and when compared to the V600E mutation, the placement of the activation loop is quite different. But looking closely at some of our predicted structures we observe that occasionally a specific conformation in the cluster has an oddly positioned residue that does more resemble the mutated counterpart instead of its cluster representants. For example prediction 6 in cluster C(wt) in orange on [figure 19] has the red 600 residue in an ‘active like’ conformation resembling more at predictions in the C(V600E) cluster than its own cluster. Curious about the prevalence of such conformation errors we decided to run the wildtype prediction over 100 times and inspect for each result the exact position of the residue 600 backbone Cα**.** [Figure 21] shows the results on a graph where the 100 models are plotted in grey and where 2 reference models for the wildtype (blue) and mutated (red). The prevalence of conformations where the activation loop resembles the mutated structure more than the wildtype is around 7%. Those results might be the consequence of the AF3 model training data containing examples of both BRAF isoforms skewing the results toward training data examples. It is also possible that AF3 is able to predict small confirmation changes but there is a small percentage of chance the algorithm predicts the wrong conformation. Another possibility is that more flexible regions like zones that are regularly mutated and not so well conserved in evolution are harder to predict using AF3 models. Thus, one question that came to mind was, how to quantify the protein regions with the most variability using C(wt) wild type (cluster containing prediction of BRAF wild type and dabrafenib) and C(V600E) containing the predictions for BRAF V600E. A common way of quantifiably measuring small placement differences of atoms in 3D structures is by using the Root Mean Square Difference (RMSD). More specifically, we computed, as illustrated in the Methods section, the average per-residue RMSD with respect to a reference model by

averaging the RMSDs of all the models. We apply this procedure once on each cluster C(wt) and C(V600E) then take the absolute value of the difference between C(wt) and C(V600E). The results are shown in [figure 20].

We observe that only a few residues, located in specific regions of the protein, show the highest variability in AF3 predictions. These residues are found at the interface between BRAF and MEK, specifically at sites 464-469 in the BRAF sequence, near the mutation site residues 595–615, and predominantly near flexible loops or hinge regions. These regions could be potentially linked to the structural effects of the mutations on the protein, suggesting that V600E leads to BRAF activation through a stronger binding with MEK.

The same clustering method and analysis was performed using a confidence metric of AF3. The pLDDT metric indicated the model's confidence for each residue in the final prediction. We can use this confidence metric again to assess differences between the C(wt) and C(V600E) clusters in a similar way we did with the RMSD by averaging pLDDT over all residues in our cluster and then comparing between the wild type and mutated cluster. Results retrieved using this metric are very similar to the RMSD results and show the same trend where residues involved at interfaces between 2 proteins or close to the active site have more variability in their prediction results using models like AF3.

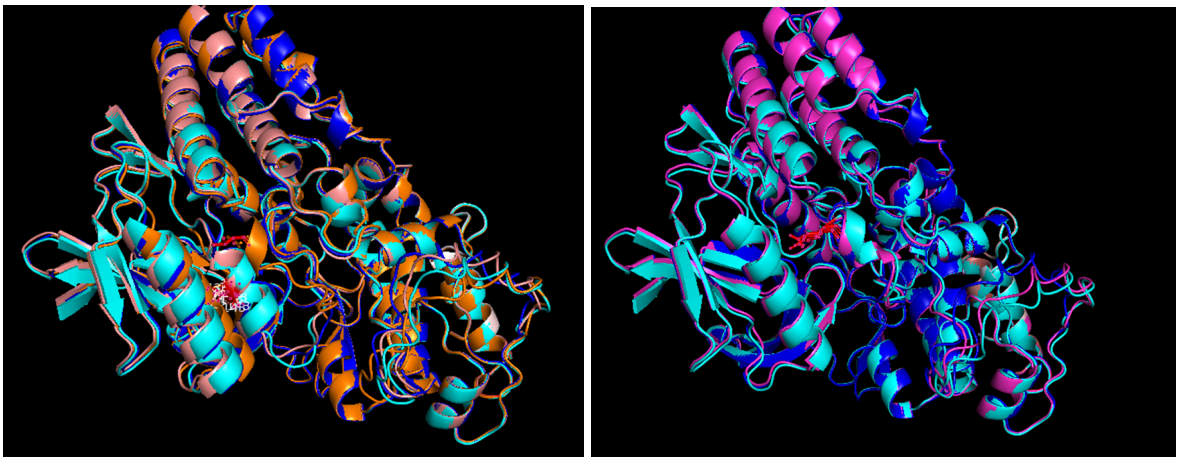


Figure 19: *AF3 predicted conformation cluster. RED residue represents residue 600 in both clusters.Model 8 in wildtype predictions has structure similar to mut_V600E models predictions. Also model 9 in mut_V600E is very similar to wildtype cluster*.

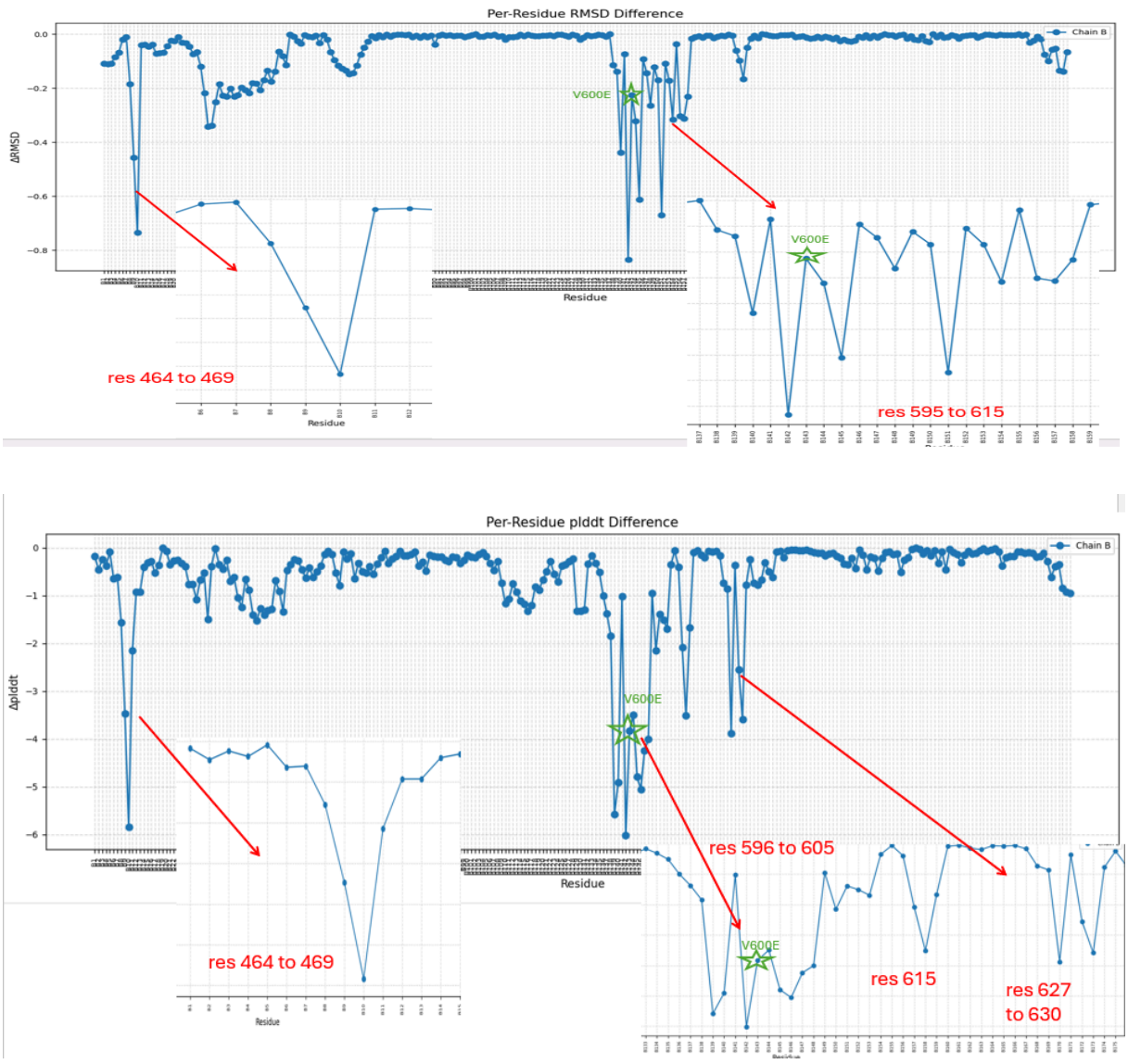


Figure 20: *(a) Top panel shows the absolute per-residue RMSD differences between the wild-type and V600E-mutant BRAF cluster, highlighting regions of high structural deviation near the mutation site (residues 595–615) and interface sites (e.g., 464–469). (b) Bottom panel shows the per-residue pLDDT confidence scores revealing localized changes in prediction confidence around the V600E site (residues 596–605, 615) and other flexible regions (464–469, 627–630).*

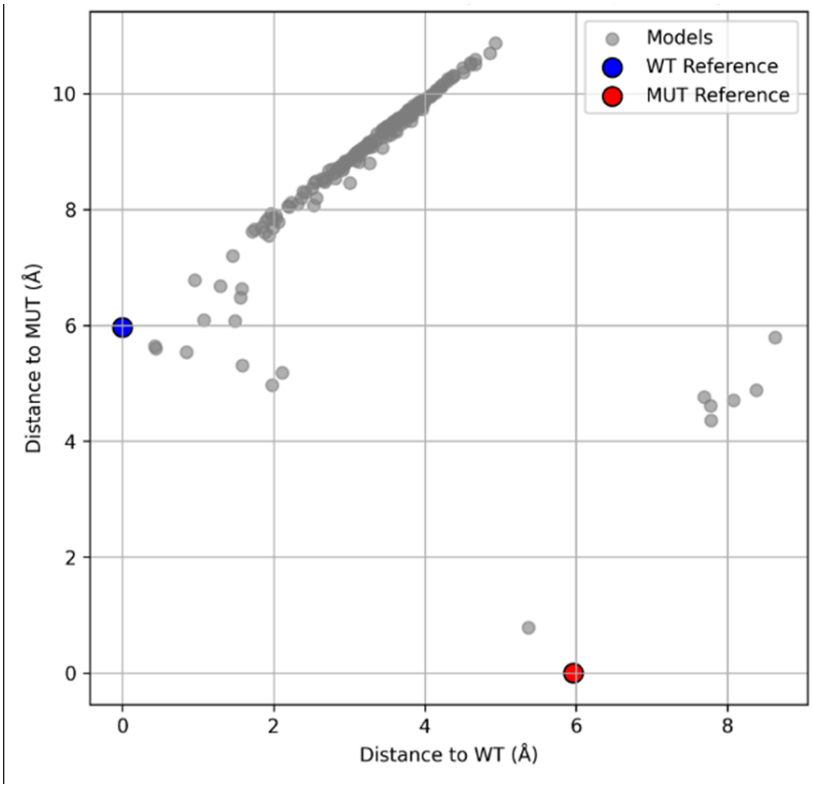


Figure 21: *Euclidean distance in 3D space between the Cα atom (alpha carbon) of residue 600 in each model vs. the aligned reference structures. Each grey point represents one model, plotted by its distance to the WT reference (x-axis) and to the MUT reference (y-axis).*

4.3.3 Protein-Protein and Protein-ligand Binding affinity predictions

In this section, we describe how we used online tools to define precise binding affinity and complex stability values using prediction via reference pdb structures and simulated AF3 model predictions. The 3 main tools that have been tested here are not representative of the whole field but are state-of-the art predictors. The first 2 tools are named BeAtMuSiC and PremPLI and are prediction tools for prediction of protein-protein and protein-ligand binding affinity change (△△G), respectively, upon mutations. Unfortunately, those 2 predictors were not reliable tools to predict binding affinity between BRAF and target molecules for different reasons that we will describe in more detail. Finally, we also used Boltz 2 predictor that is a tool constructed on the AF3 infrastructure to predict binding affinity parameters.

Mutation	ΔΔGbind – AF3 model wildtype
V600E / E600V	**+0.90**
K601E	0.02
K601N	0.04
K601T	-0.11
L597S	0.51
L597Q	0.44
L597V	0.33
G469A	**1.01**
G469V	**0.97**
G469R	**1.00**
G469S	**1.21**
G464V	0.45
G464E	0.88
T599I	0.37

Table 4: *BRAF-MEK binding affinity results for BRAF common variants based on BeAtMuSiC on Alpha fold 3 predicted models. Negative* △△G is indicated by green while positive △△G are in red.

The first idea was to assess the effect of different mutations identified in the literature on BRAF. The first tool we used for this was developed by our lab and is called BeAtMuSiC [40]. The main use of the program is to predict the thermodynamic stability of a complex of proteins, in our case the MEK-BRAF protein complex. The assumption is that that genetic variant of BRAF that lead to increase thermodynamic stability (△△G) for the BRAF-MEK complex would lead to a more biologically active complex resulting in higher chances of cancer development. To test this idea, we used the top alpha fold generated models from the previous section. This comparison was made to evaluate how BeAtMuSiC would differ between nearly identical BRAF models. Table 4 displays the results of BeAtMuSiC when it is launched on AF3 predicted structures of BRAF variants. Unfortunately, the results were underwhelming as no consistency with the literature was found. Mutations like V600E that should result in a conformational change facilitating their kinase activation of the MEK1 target and thus enhance the binding affinity were not predicted as important but as weakening the affinity which is contrary to what we expect in vivo. However, this

reflects one of the major problems with current state-of-the-art predictors: while they perform relatively well in assessing loss-of-function mutations, they struggle to predict gain-of-function mutations effectively.

Another tool was used to evaluate the folding free energy (△△G) between the BRAF protein and the target drug dabrafenib. The hypothesis is that mutations close to the binding site of dabrafenib can affect its binding and can lead to cancer escape from the drug. More in detail, we used PremPLI, which is a machine learning model for predicting the effects of missense mutations on protein-ligand interactions [41]. Results are shown in Table 5 and show that also in this case free energy of the complex is systematically higher, except for the E600V variant, it seems to be more stable than the V600E variant leading to lower △△G. Another problem that deserves further investigation is the lack of consistency when using different input structures in both tools. Many predictions made with one PDB structure were different when using a similar crystal structure or an Alpha Fold-predicted model, indicating that there is a source of noise in the structure and in the prediction models.

Mutation	△△Gbind for BRAF + dabrafenib (AF3-model)
E600V	**-0.95**
V600E	**+0.5**
K601E	**+0.82**
T599I	**+0.6**
G464E	**+0.9**
L597V	**+0.62**

Table 5: *BRAF-dabrafenib binding affinity results for BRAF common variants predicted by PremPLI on best Alpha fold 3 predicted model. Negative* △△G is indicated by green while positive △△G are in red.

The last bioinformatics tool that was used in this work is called Boltz 2 that can predict log(IC50) value of a drug-target complex making comparisons with the GDSC database much easier. To ensure ourselves of the high quality of Boltz 2 results, a reliable way to assess the model quality was to visualize and compare using Pymol the structure predictions from both programs as can be seen on [figure 22]. We observe that both predicted structures are close to identical, indicating a good fidelity to Alpha Fold 3 structures for Boltz 2 predictions. The main idea in this calculation is to compare variants of BRAF when interacting with different drug compounds and to extract via Boltz 2 the binding affinities of them. Therefore, we started by extracting the different BRAF variants from all cancer cell lines in the GDSC database. Insertions, deletions or other rearrangements were discarded from the analysis but could be added in further analysis. This resulted in a simple pipeline where all BRAF single nucleotide variants were evaluated against dabrafenib in a manner that could easily be automated to test on other drug-gene combinations. Table 6 shows the results of Boltz 2. Unfortunately, when analyzing the results the Spearman correlation between the Boltz 2 predicted BRAF-dabrafenib affinity in the given cancer cell lines and the measured effect of the drug in the cancer cell line was poor and equal to -0.11. This indicates that Boltz 2 predictions were not precise enough to distinguish a significant difference in binding affinity to the drug when the only changes to the molecule were single amino acid substitution. For example, dabrafenib is a mutant BRAF-specific inhibitor with an IC50 of about 0.7 nM, which is one order of magnitude stronger than the inhibition of wild-type BRAF. The log(IC50) should thus result in a factor of two differences, while Boltz 2 yields very close values, with slightly stronger inhibition for BRAF wild-type than for BRAF V600E [figure 23]. This result was underwhelming because this might signify that Boltz 2 is not accurate enough to evaluate the effect of mutation on the drug-target affinity.

One possible reason for this result is that we isolated only the MEK-binding domain of BRAF instead of using the full protein structure. This approach was chosen to improve Alpha Fold 3 and Boltz 2 confidence, but it might have missed important conformational details or interactions. However, testing with the entire BRAF sequence did not improve the results and actually made them more inconsistent with the literature. Another possible problem was that the molecular environment surrounding the BRAF kinase activation is too complex to be modeled in isolation.

This is caused by the fact that BRAF activation in the cell is a complex process that is overseen by many helper molecules like the chaperone 14-3-3. Also, BRAF tends to form homo or heterodimers in its wildtype form where it is able to autoactivate when mutated. All those molecular mechanics were not captured in our test, but they could be included in a following analysis.

We then searched for another drug/target pair with much simpler interactions where a single mutation causes a drastic change in binding affinity due to a conformational change of the target protein. Such an example was found after some research, the *FLT3* gene codes for the identically named protein and is important for the development of blood vessels. Mutations like D835H or D835Y have a significant effect on the 3-dimensional structure that drastically modify its affinity for quizartinib, a type II TKI used in acute myeloid leukemia (AML). In 2013, *Lin and al*. [44] showed that the D835H variant conferred ~26-fold resistance to quizartinib compared to the FLT3-ITD mutation alone. While the D835Y variant has resistance levels exceeding 100-fold. Finally, the D835E variant has the least resistance to quizartinib (<20 fold). Table 7 shows the results when reproducing those affinity values using Boltz 2. We observe that the predicted binding affinities here are also inconsistent with the literature, with the exception of D835E. However, even in this case, the difference is much smaller than what is reported in experimental studies. This indicates once again that Boltz 2 might not be precise enough to capture small conformational changes and their relation to binding affinity.

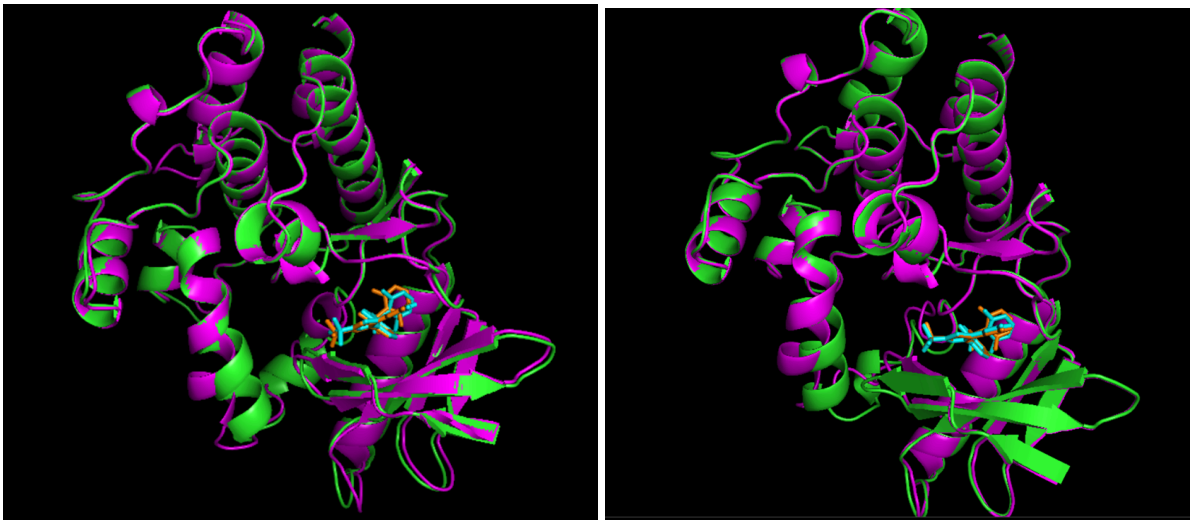


Figure 22: Structural alignment of Boltz 2 (magenta) and Alpha Fold 3 (green) models for BRAF wild-type (left) and V600E mutant (right), highlighting local differences near the active site and mutation site. Dabrafenib is shown in orange for Alpha Fold 3 and cyan for Boltz 2.

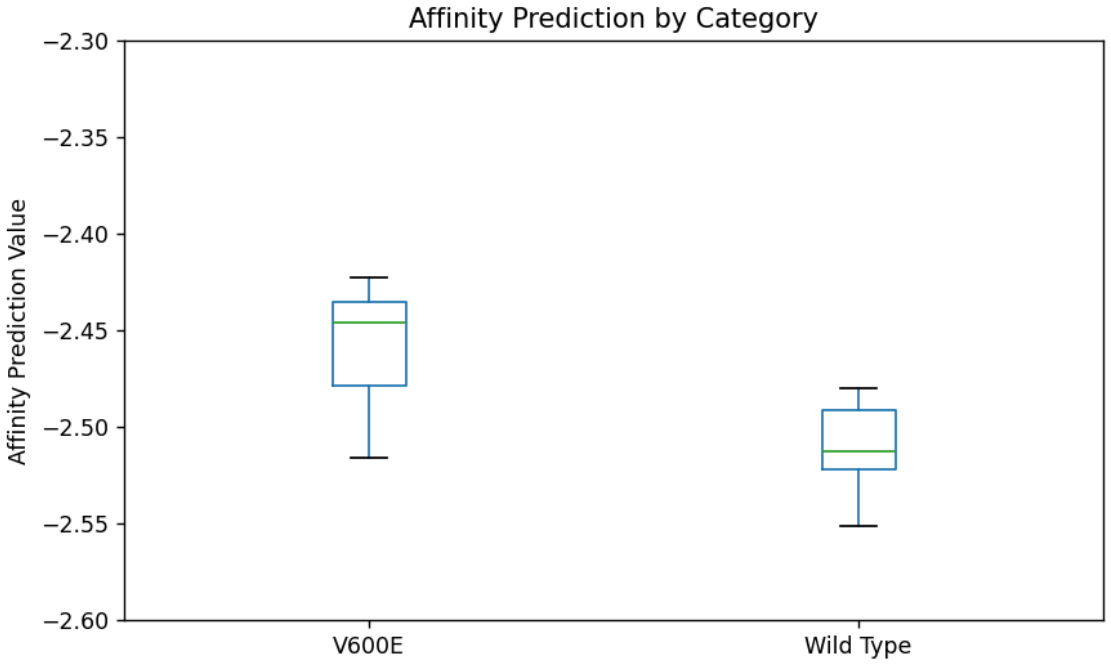


Figure 23: Boxplot showing 20 predictions realized using boltz 2 affinity prediction module using identical inputs. Left boxplot represents 10 BRAF V600E affinity prediction with dabrafenib, Right boxplot represents 10 BRAF wild type affinity prediction with dabrafenib.

BRAF Mutation	Predicted affinity log(IC50)	GDSC affinity log(IC50)	BRAF Mutation	Predicted affinity log(IC50)	GDSC affinity log(IC50)
Wild Type	-2.59	4.17	G596R	-2.53	3.48
V600E	-2.48	-0.60	K499T	-2.48	/
V600M	-2.53	0.15	M650I	-2.54	/
D594G	-2.61	/	K601N	-2.44	1.04
G464E	-2.26	1.25	L597R	-2.47	4.96
G464V	-2.13	3.90	L597V	-2.50	2.76
G466V	-2.51	3.70	N581Y	-2.56	3.81
G469A	-2.47	3.37	V641I	-2.47	/
G469V	-2.43	1.88	T529A	-2.48	5.14

Table 6: Table showing the affinity predictions from the Boltz 2 program for different variants of the BRAF protein. Results showing that most variants have a similar sensitivity to dabrafenib. Experimental GDSC results are shown for comparison. Spearman Correlation between predicted and experimental affinities is -0.11.

FLT Mutation	Predicted affinity log(IC50)
Wild Type	-1.419
D835Y	-1.348
D835H	-1.335
D835E	-1.611

Table 7: Table showing the affinity predictions from the Boltz 2 program for different variants of the FLT3 protein. Results show that the D835E variant is the most sensitive while D835Y and D835H are more resistant to quizartinib.

4.3.4 Building a structural predictor for drug-target identification

In the previous section we tried to see how mutating the driver gene BRAF in order to predict its binding affinity with a specific drug has not been successful using Boltz 2 and PremPLI affinity predictions. In this section we tackle another approach that consists in testing a wider range of drugs on the 2 main mutational variants of BRAF (wildtype and V600E). This is a slightly simpler problem and can be understood as a form of in-silico drug discovery procedure aiming at finding potential hits for a specific target by using a large database of known drug compounds and selecting potential binders before more advanced biopharmaceutical screening and modification of the candidate molecules. The main strength of this approach is that it is easily scalable and adaptable to other targets for future analysis. Here, we get the results for both BRAF variants. Figure 24 shows a schematic version of the pipeline as an illustration of the whole procedure. Final results are stored in csv format and available in the supplementary material. The next section dives deeper into the core results of the analysis.

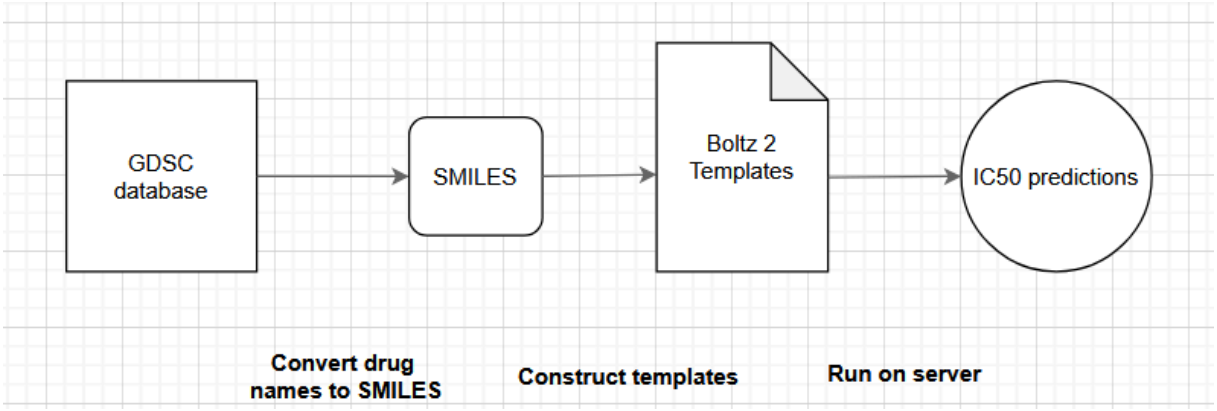


Figure 24: Schematic representation of the boltz 2 drug sensitivity analysis pipeline.

4.3.5 Correlations with IC50

We previously used the binding affinity predictor Boltz 2 to predict the affinity between BRAF protein and all drug compounds present in GDSC. We can then test our predictors against the original GDSC data and figure out how accurate the prediction pipeline is to detect potential sensible drug compounds for the protein targets. We started by determining IC50 values for all the drugs against BRAF then we plotted those values on a horizontal axis to reproduce the volcano plot results from the GDSC webserver. Figure 25 shows a comparison between the original experimental data from GDSC (fig 25a) against the reproduction realized using Boltz 2 predictor (fig 25b). A major observation is that sensitive drugs such as dabrafenib or (5Z)-7-Oxozeaenol are detected by the predictor. Figure 26a shows the Spearman correlation between the experimental data and predicted data when looking at the binding affinity against multiple drugs. The global Spearman correlation is 0.248. This weak correlation shows that there is a small overlap between our prediction method and experimental cell line level cancer data which is encouraging but not good enough to make robust predictions in drug discovery pipelines. On figure 26b we see the correlation between BRAF V600E affinity predictions compared to BRAF wild type using Boltz 2 on the GDSC drug panel. High Spearman correlation (0.984) between those predicted affinities reaffirms that Boltz 2 lacks the sufficient resolution to accurately predict single point mutations and their effect on protein conformations.

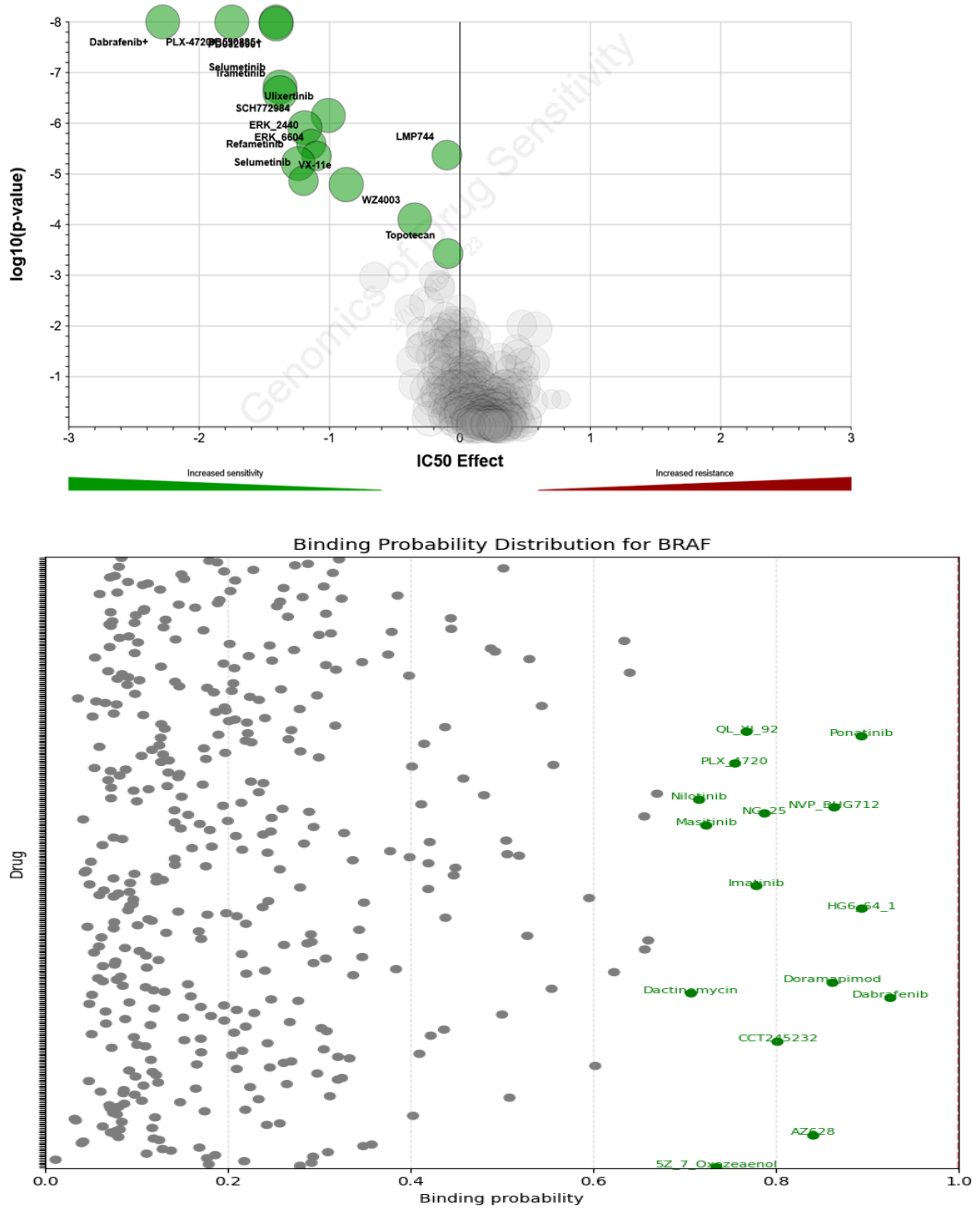


Figure 25: *(a)Volcano plot from the GDSC database showing drugs with significant sensitivity effects on BRAF-mutant cell lines, based on IC50 values and statistical significance. (b) Binding probability distribution generated using the Boltz 2 predictor, which estimates the affinity between BRAF and all compounds in the GDSC dataset. Highly sensitive drugs such as dabrafenib and 5-fluorouracil are correctly identified by the predictor.*

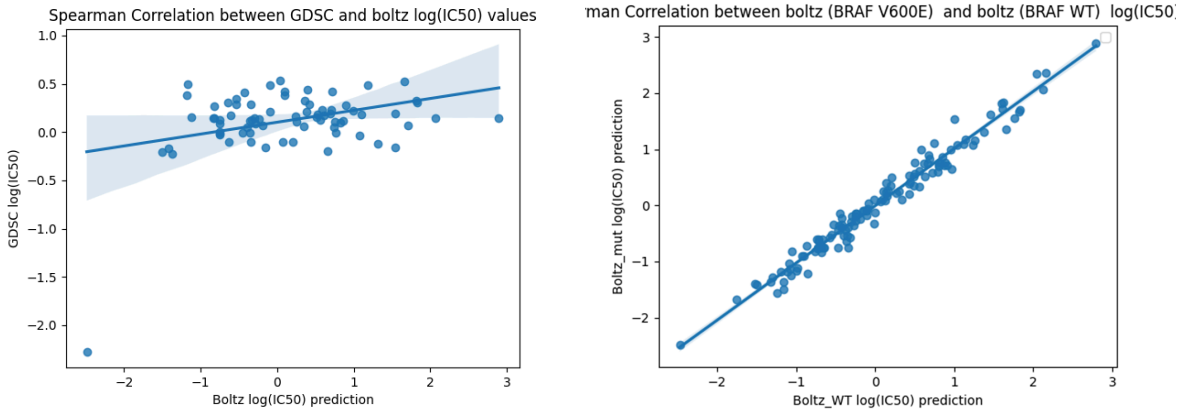


Figure 26: *(a) Spearman correlation between Boltz 2 predictions and GDSC log(IC50) values, showing moderate agreement of 0.248. (b) High correlation (0.984) between Boltz 2 scores computed from BRAF wild-type and V600E structural models, indicating consistency across mutation contexts.*

4.4 Expression data

Cancer drug resistance is a very complex mechanism, and drug-target affinity is just one component of the bigger picture. As mentioned in the introduction, many other biological mechanisms are used by cancer cells to escape drugs. Here, we aim to acquire new features that might better explain these biological processes. Typically, we want to extract strong biologically relevant data that could help to uncover patterns that cannot be directly related to mutational or structural data, for example, the expression of genes in the cells. The study of gene expression is very important in understanding the dynamic state of a cell, as it indicates which genes are actively being transcribed and potentially translated into functional proteins. Contrary to more “static” features, such as mutational signature, gene expression levels give indications on the regulation mechanisms of the cell as well as the external environment influences on drug intake by the cancer cells. The impact of point mutations can also be observed in gene expression, exemplified by the well-known TERT promoter mutations that enhance TERT expression [45].

In the case of GDSC, the gene expression levels for our cell lines were determined and are available for free download on the cell model passport website. In this section we integrated the gene expression data for the GDSC cell lines by looking at major differences between *BRAF* wild type and *BRAF* V600E mutated cell lines. For

each cell line in the dataset, it was determined if it contained the V600E mutation or not using the same filtering as in section 3.2. Non-V600E mutations were filtered out of the data keeping a dataset of 2 subsets representing the V600E and the wild type cell lines. Table 8 shows the top 10 gene expressions that are the most impacted by the mutation as measured by the log fold change of their expression between the *BRAF* V600E and wild type cancer cell lines. We observe that many of those genes correspond to IGK (Immunoglobulin Kappa) variations, suggesting an immune-related signature related to the *BRAF* V600E mutation. This indicates that the mutation may not only affect cancer cell growth but could also influence immune responses, potentially contributing to immune evasion or altered immune system interactions.

Gene	Log Fold Change	P-value Adjusted
IGKJ1	27.495	6.079406e-34
IGLL5	25.655	1.855249e-37
LILRB1	25.293	1.751356e-51
MIR205	24.86	2.402792e-82
VPREB1	24.196	1.021461e-25
IGKJ4	23.854	1.912932e-27
IGKJ5	23.791	4.251086e-27
IGKV3-20	13.05	1.120346e-05
IGKC	12.431	1.002806e-19
KRT5	12.358	3.113857e-38
MS4A1	12.055	1.479706e-14

Table 8: *top 10 differentially expressed genes between BRAF V600E-mutant and wild type cancer cell lines ranked by log fold change.*

The next step in this analysis was to utilize the expression of these 10 selected features to build a linear model that integrates both structural predictions (the binding affinity between the drug and its target from the Boltz 2 model) and expression data to predict drug sensitivity. This combined model was evaluated against observed GDSC IC50 values. Results are shown in figure 27. The integration of expression features returned a small but consistent improvement in predictive performance, with the Spearman correlation increasing from the baseline to 0.255. This result supports the idea that transcriptomic features can give indications, even if weak, on additional biological variance related to drug response.

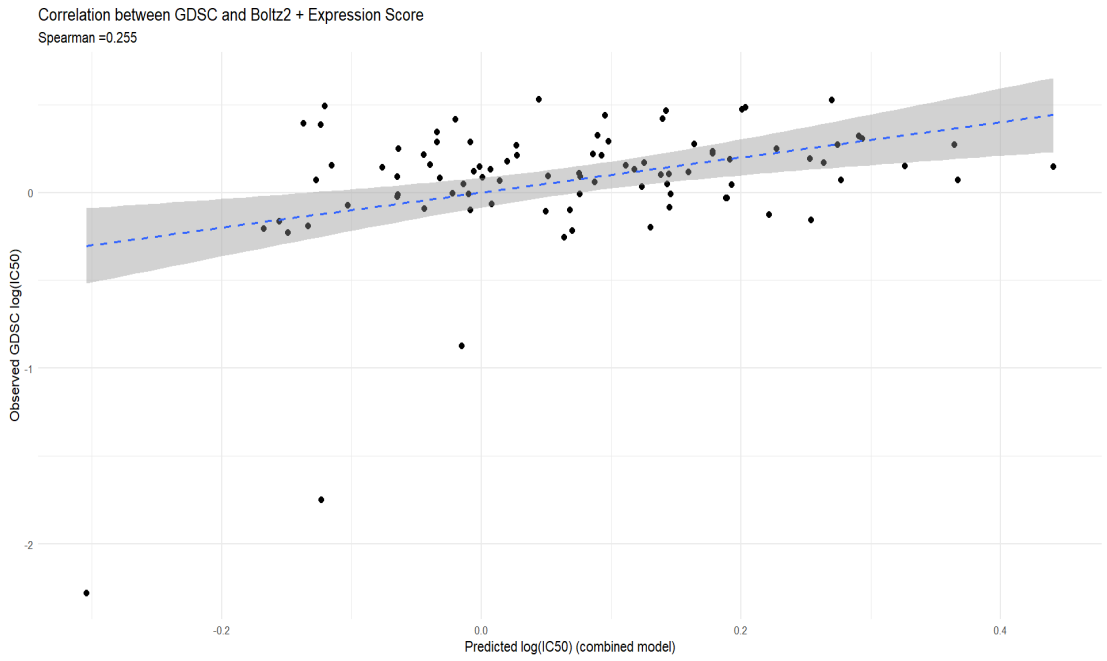


Figure 27: *Spearman correlation between our predictor (Boltz 2 predictions + top 10 differentially expressed genes) and GDSC log(IC50) values, showing slight improvement in spearman correlation to 0.255.*

Chapter 5 – Discussion and Perspectives

5.1 Structural bioinformatics for drug response prediction

Structural bioinformatic approaches in drug sensitivity prediction are still underexplored but offer promising results. Structural data about the interface between a drug and its protein target can provide biologically relevant information that can be easily generalizable. However, it is challenging to interpret the relationships between protein structure, drug-binding and function, especially in presence of complex protein dynamics or conformational flexibility. Compared to popular in silico techniques like molecular dynamics simulations or docking, our approach involving deep learning methods (e.g., Alpha Fold 3, Boltz 2), was computationally cheaper and more scalable. However, despite the claims in the original papers, it might lack the statistical strength required to demonstrate reliable drug sensitivity prediction.

5.2 Strengths and limitations of the approach

Describing methodological and biological strengths of this approach: the pipeline is modular and scalable and allows rapid testing of other gene-drug pairs with few requirements for manual intervention. The use of open-access tools and resources (Alpha Fold, GDSC, Boltz 2) makes it reproducible and adaptable, even for other genes than the BRAF example. The combination of structure, expression, and sensitivity data provides a wider view of drug dynamics giving a more complete understanding of cancer pharmacogenomics.

On the other hand, the approach has several limitations that must be taken into account. First and most importantly, deep learning prediction methods seem to underperform compared to the claims in the original papers. However, additional factors could be involved. For example, the GDSC drug response data are often noisy and not always complete. A major difference between GDSC responses and

our predictor is that they operate at two distinct levels. While the GDSC measures cell lines responses, our approach focuses on the molecular-scale interface and specific protein-drug residue interactions, ignoring many other biological factors that can influence the final result like immune responses, secondary mutations in other pathways, etc… Moreover, biological interpretation of Boltz 2 scores is still evolving. These scores approximate energy landscapes and do not account for dynamics or post-translational modifications.

Integrating transcriptomic data into a simplified score to aggregate the data is certainly an interesting way that we explored. However, the resulting updated model's prediction strength is only marginally better than the structure-based model. This step could be enhanced by incorporating pathway-level analyses or machine learning–based feature selection in future iterations.

Despite these limitations, the correlation between structural predictions and known drug responses for key malignant mutations shows that these tools can generate new hypotheses for drug discovery at early stages but might require more reliable methods for rigorous prediction of biological activity.

5.3 Future directions and perspectives

This work has multiple possibilities for improvement. First, applying this approach to other driver genes or tumor suppressing genes for different drug classes would help assess its generalizability. For instance, exploring EGFR, KRAS, or PI3K pathways could reveal new insights that further validate or challenge this method. Another idea would be to add more multi-omics features and integrate proteomics, epigenomics, or even metabolomics to improve the accuracy of the model and capture more relevant biological clues. This could help resolve situations where expression or structure alone fail to explain drug effects. Third, the pipeline could be applied to other pharmacogenomic datasets such as PRISM or CCLE to validate our results and identify differences between datasets. This could help in strengthening our findings by reproducing similar results on unseen data. Finally, there is clearly a need for

better tools and metrics for the effects of structural perturbations caused by mutations. Developing machine learning models trained specifically on binding affinity shifts when mutations are introduced in the target protein could be an idea to improve those results but it requires high quality affinity data that are not always available. Additionally, a multitude of new tools have been developed more recently to assess protein-ligand binding using structure or ML models, but we did not have the time to test those.

In summary, this thesis is the first attempt to construct a scalable, multi-layered computational framework to explore drug sensitivity prediction, highlighting both the potential and current limitations of structural methods in pharmacogenomics. Predicting drug sensitivity in cancer remains thus an open challenge that might highly benefit from a more accurate method to predict the impact of mutations on protein-ligand binding, biologically relevant multi-omics integration and a deeper understanding of the cellular ecosystem. With further refinement, our approaches based on the integration of protein structural data with multi-omics information could help in identifying new resistance mechanisms, developing and optimising drugs, and ultimately advancing personalized cancer therapy.

Chapter 6 – Conclusion

Using structural analysis to discover molecular mechanisms that play a role in cancer drug sensitivity is a complex task. Additional high quality multi-omic data is required to obtain a realistic picture of the many intervening components as well as a strong binding affinity predictor or state-of-the-art docking techniques that accurately predict the 3D structures of the interest molecules and their precise interaction process. It requires those tools to be precise enough in their predictions to detect subtle changes in protein conformation due to mutations and how they impact their binding strength. In this paper, we explored a new approach, taking advantage of the recent explosion in artificial intelligence and especially deep learning methods like Alpha fold of Google Deepmind. The latest version of Alpha Fold 3 is able to accurately predict molecular interactions with high resolution for a wide range of molecules. To predict the binding affinity we used another deep learning tool called Boltz 2, based on alpha fold, it adds specific modules that specifically evaluate the strength of a binder with its target. We analysed two other tools called BeAtMuSiC and PremPLI with the same objective as Boltz 2 to assess the thermodynamic stability of the target protein drug complex. To do this analysis we started by doing a profound research in the literature to find good candidates for potent oncogenes. We finally settled on the BRAF gene because it is well-described in the literature and is commonly mutated in different types of cancer like metastatic colorectal cancer or melanoma. It was also described how the V600E mutation induced a conformational change in the protein that would lead to overactivation of the MAPK/ERK pathway. The results in this paper show that even if state-of-the-art deep learning tools make accurate structural predictions in simple mono or dimolecular environments, the resolution to detect small genetic modification and their impact on protein-ligand binding is insufficient. This is due to inherent limitations of recent machine learning techniques often prone to hallucination or overfitting of the training data. Then, we used the Boltz 2 algorithm to try to predict the sensitivity of several drugs. These Boltzmann scores were correlated with IC50 values from the Genomic of Drug Sensitivity in Cancer (GDSC) dataset, revealing moderate (0.248) but biologically coherent correlations for drug response results. Finally, we integrated transcriptomic data using gene expression to assess whether combining structural and expression-based signals improved the

predictions. While the expression features were not very helpful across all drugs, it added value in certain cases, suggesting that multi-modal approaches capture complementary aspects of drug response. The correlation to GDSC data improved to 0.255 with this approach. In summary, these findings show that the current state of the art structural prediction deep learning algorithms like Alpha Fold 3 or Boltz 2 are unable in isolation to reliably predict the sensitivity of drug compounds in relation to their protein target due to the biological complexity of cancer resistance and sensitivity mechanisms and the inherent limitations of those techniques. The addition of other biologically relevant features like expression profiles for cell lines should have a small positive impact on the final prediction accuracy but due to a lack of time it was not profoundly analysed in this thesis. Ultimately, predicting drug sensitivity in cancer remains an open challenge that might highly benefit from accurate structure prediction, biologically relevant multi-omic integration and a deeper understanding of the cellular chemical ecosystem.

References

1. *Lifetime Risk of Developing or Dying From Cancer [Internet]. Available from: https://www.cancer.org/cancer/risk-prevention/understanding-cancer-risk/lifetime-probability-of-developing-or-dying-from-cancer.html*

2. *Brown JS, Amend SR, Austin RH, Gatenby RA, Hammarlund EU, Pienta KJ. Updating the Definition of Cancer. Mol Cancer Res. 2023 Nov 1;21(11):1142–7.*

3. *Hanahan D. Hallmarks of Cancer: New Dimensions. Cancer Discovery. 2022 Jan 12;12(1):31–46.*

4. *McFarland CD, Mirny LA, Korolev KS. Tug-of-war between driver and passenger mutations in cancer and other adaptive processes. Proceedings of the National Academy of Sciences. 2014 Oct 21;111(42):15138–43.*

5. *Liu CH, Lai YL, Shen PC, Liu HC, Tsai MH, Wang YD, et al. DriverDBv4: a multi-omics integration database for cancer driver gene research. Nucleic Acids Res. 2024 Jan 5;52(D1):D1246–52.*

6. *TCGA - PanCanAtlas Publications | NCI Genomic Data Commons [Internet]. [cited 2025 Aug 10]. Available from: https://gdc.cancer.gov/about-data/publications/pancanatlas*

7. *Muiños F, Martínez-Jiménez F, Pich O, Gonzalez-Perez A, Lopez-Bigas N. In silico saturation mutagenesis of cancer genes. Nature. 2021 Aug;596(7872):428–32.*

8. *The Cancer Genome Atlas Program (TCGA) | [Internet]. [cited 2025 Aug 10]. Available from: https://med.und.edu/research/transcend/_files/pdfs/berdc_resource_pdfs/april_2024_sl_tcga.pdf*

9. *Cancer Research UK [Internet]. 2015 [cited 2025 Jun 18]. Cancer survival for common cancers. Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/survival/common-cancers-compared*

10. *2022 Cancer Facts & Figures Cancer | Cancer Death Rate Drops [Internet]. [cited 2025 Aug 10]. Available from: https://www.cancer.org/research/acs-research-news/facts-and-figures-2022.html*

11. *Lei ZN, Tian Q, Teng QX, Wurpel JND, Zeng L, Pan Y, et al. Understanding and targeting resistance mechanisms in cancer. MedComm (2020). 2023 Jun;4(3):e265.*

12. *Mansoori B, Mohammadi A, Davudian S, Shirjang S, Baradaran B. The Different Mechanisms of Cancer Drug Resistance: A Brief Review. Adv Pharm Bull. 2017 Sep;7(3):339–48.*

13. *Drug Uptake - an overview | ScienceDirect Topics [Internet]. [cited 2025 Aug 10]. Available from: https://www.sciencedirect.com/topics/immunology-and-microbiology/drug-uptake*

14. *Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J., & Bolton, E. E. (2025). PubChem 2025 update. Nucleic Acids Res., 53(D1), D1516–D1525.*

15. *H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank (2000) Nucleic Acids Research 28: 235-242.*

16. *Ahmad S, da Costa Gonzales L J, Bowler-Barnett E H, Rice D L, Kim M, Wijerathne S, Luciani A, Kandasaamy S, Luo J, Watkins X, Turner E, Martin M J, UniProt Consortium The UniProt website API: facilitating programmatic access to protein knowledge Nucleic Acids Research, (2025)*

17. *Kanehisa M, Furumichi M, Sato Y, Matsuura Y, Ishiguro-Watanabe M. KEGG: biological systems database as a model of the real world. Nucleic Acids Research. 2025 Jan 6;53(D1):D672–7.*

18. *Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012 Mar 28;483(7391):603–7.*

19. *Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. 2013 Jan 1;(41):D955-61.*

20. *Marin JJG, Herraez E, Lozano E, Macias RIR, Briz O. Models for Understanding Resistance to Chemotherapy in Liver Cancer. Cancers. 2019 Nov;11(11):1677.*

21. *Castellani G, Buccarelli M, Arasi MB, Rossi S, Pisanu ME, Bellenghi M, et al. BRAF Mutations in Melanoma: Biological Aspects, Therapeutic Implications, and Circulating Biomarkers. Cancers. 2023 Jan;15(16):4026.*

22. *Śmiech M, Leszczyński P, Kono H, Wardell C, Taniguchi H. Emerging BRAF Mutations in Cancer Progression and Their Possible Effects on Transcriptional Networks. Genes. 2020 Nov;11(11):1342.*

23. *Durrant DE, Morrison DK. Targeting the Raf kinases in human cancer: the Raf dimer dilemma. Br J Cancer. 2018 Jan;118(1):3–8.*

24. *Sun Q, Wang W. Structures of BRAF–MEK1–14-3-3 sheds light on drug discovery. Sig Transduct Target Ther. 2019 Dec 13;4(1):1–2.*

25. *Xu T, Wang X, Wang Z, Deng T, Qi C, Liu D, et al. Molecular mechanisms underlying the resistance of BRAF V600E-mutant metastatic colorectal cancer to EGFR/BRAF inhibitors. Ther Adv Med Oncol. 2022;14:17588359221105022.*

26. *Oscier D, Stamatopoulos K, Mirandari A, Strefford J. The Genomics of Hairy Cell Leukaemia and Splenic Diffuse Red Pulp Lymphoma. Cancers (Basel). 2022 Jan 29;14(3):697.*

27. *Maloney RC, Zhang M, Liu Y, Jang H, Nussinov R. The mechanism of activation of MEK1 by B-Raf and KSR1. Cell Mol Life Sci. 2022 May 4;79(5):281.*

28. *Yue Z, Zhang W, Lu Y, Yang Q, Ding Q, Xia J, et al. Prediction of cancer cell sensitivity to natural products based on genomic and chemical properties. PeerJ. 2015;3:e1425.*

29. *Naulaerts S, Dang CC, Ballester PJ. Precision and recall oncology: combining multiple gene mutations for improved identification of drug-sensitive tumours. Oncotarget. 2017 Nov 14;8(57):97025–40.*

30. *Qin Y, Conley AP, Grimm EA, Roszik J. A tool for discovering drug sensitivity and gene expression associations in cancer cells. PLoS One. 2017;12(4):e0176763.*

31. *Fei Zhang, Wang M, Xi J, Yang J, Li A. A novel heterogeneous network-based method for drug response prediction in cancer cell lines. Sci Rep. 2018 Feb 20;8(1):3355.*

32. *Ha S, Park J, Jo K. Comparative analysis of regression algorithms for drug response prediction using GDSC dataset. BMC Res Notes. 2025 Jan 13;18(1):10.*

33. *Malhotra Y, John J, Yadav D, Sharma D, Vanshika, Rawal K, et al. Advancements in protein structure prediction: A comparative overview of AlphaFold and its derivatives. Computers in Biology and Medicine. 2025 Apr 1;188:109842.*

34. *Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024 Jun;630(8016):493–500.*

35. *Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021 Aug;596(7873):583–9.*

36. *Passaro S, Corso G, Wohlwend J, Reveiz M, Thaler S, Somnath VR, et al. Towards Accurate and Efficient Binding Affinity Prediction.*

37. *The PyMOL Molecular Graphics System, Version 1.74 Schrödinger, LLC. | pymol.org [Internet]. [cited 2025 Aug 10]. Available from: https://www.pymol.org/*

38. *Drug Download Page - Cancerrxgene - Genomics of Drug Sensitivity in Cancer [Internet]. [cited 2025 Aug 10]. Available from: https://www.cancerrxgene.org/downloads/genetic_features?mutation=variant*

39. *van der Meer D, Barthorpe S, Yang W, Lightfoot H, Hall C, Gilbert J, et al. Cell Model Passports—a hub for clinical, genetic and functional datasets of preclinical cancer models. Nucleic Acids Res. 2019 Jan 8;47(D1):D923–9.*

40. *Dehouck Y, Kwasigroch JM, Rooman M, Gilis D. BeAtMuSiC: prediction of changes in protein–protein binding affinity on mutations. Nucleic Acids Research. 2013 Jul 1;41(W1):W333–9.*

41. *Sun T, Chen Y, Wen Y, Zhu Z, Li M. PremPLI: a machine learning model for predicting the effects of missense mutations on protein-ligand interactions. Commun Biol. 2021 Nov 19;4(1):1311.*

42. *Love MI, Huber W, Anders S (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." Genome Biology, (15), 550.*

43. *AlphaFold Server [Internet]. [cited 2025 Aug 10]. Available from: https://alphafoldserver.com/*

44. *Lin K, Smith CC, Salerno S, Shah NP. Mutations At The FLT3 Activation Loop D835 Residue Confer Differential Resistance To Clinically Active FLT3 Inhibitors. Blood. 2013 Nov 15;122(21):3929.*

45. *Min J, Shay JW. TERT Promoter Mutations Enhance Telomerase Activation by Long-range Chromatin Interactions. Cancer Discov. 2016 Nov;6(11):1212–4.*

Appendix A

All supplementary information including data, code and materials are available on github at the address: *https://github.com/sisi-le-bioinformaticien/Master-Thesis-2025*