UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

FINDING KEY CHARACTERISTICS OF PROMISING COMPOUNDS FOR ANTICANCER DRUG DISCOVERY

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE

By

PAULINE RIBEYRE
Norman, Oklahoma
2017

FINDING KEY CHARACTERISTICS OF PROMISING COMPOUNDS FOR ANTICANCER DRUG DISCOVERY

A THESIS APPROVED FOR THE
SCHOOL OF INDUSTRIAL AND SYSTEMS ENGINEERING

BY

_____
Dr. Charles Nicholson, Chair

_____
Dr. Ziho Kang

_____
Dr. Shima Mohebbi

Acknowledgements

I wish to express my gratitude to the Dr. Corey Clark and the Dr. James McCormick from the Southern Methodist University of Dallas, Texas, who made this work possible by providing the data I studied and by taking the time to answer my questions.

I also thank the Dr. Charles Nicholson, my advisor at the University of Oklahoma, for his supervision and valuable advice.

Table of Contents

List of Tables

List of Figures

Abstract

Multidrug resistance is the simultaneous resistance to two or more chemically unrelated therapeutics, including some therapeutics the cell has never been exposed to. It is one of the biggest obstacles to effective cancer chemotherapy treatments. Multidrug resistance can be caused by drug efflux, an otherwise useful body mechanism that prevents a too-high drug concentration in cells, by using proteins called transporters. Some chemical compounds have the ability to sensitize the cells to the drugs by disabling these transporters. The focus of this work is to find key characteristics of compounds that may disable a specific transporter, the P-glycoprotein. Three datasets listing compounds, their values for different features, and their ability to disable the transporters are provided by experts. Using the programming language R, various data analytics methods are applied to these datasets with the objective of predicting whether compounds are P-glycoprotein inhibitors or not. The main issue encountered is the fact that the most important dataset did not contain enough samples for the number of predictor variables. Ultimately, the decision tree and random forest models prove to be the most effective in predicting the compounds' ability to disable the transporter.

Chapter 1: Introduction

According to the World Cancer Report of 2014, 39% of people will be diagnosed with cancer at some point in their life and 15% of all deaths are cancer-related (McGuire, 2014). Furthermore, approximately 40% of human cancers develop resistance to chemotherapeutic drugs (Follit, Brewer, Wise, & Vogel, 2015). One of the issues with chemotherapy is that the human body has the ability to reject drugs through a process called drug efflux, using transporter proteins. The most obvious solution to this problem, which is to use more chemotherapeutic drugs, causes serious undesirable side effects for the patient. Studies carried out by the Center for Drug Discovery, Design and Delivery (CD4) of the Southern Methodist University of Dallas, Texas have shown that after turning off the transporter proteins, using chemotherapeutic drugs in small concentrations is enough to destroy the cancerous cells.

This work is built around the multidrug resistance protein, P-glycoprotein (P-gp), a transporter protein which is responsible for drug efflux. In previous studies, some compounds proved to be efficient in reversing the multidrug resistance caused by P-gp and restoring the cancerous cells' sensitivity to the chemotherapeutic drugs. Used at the concentration where they reverse multidrug resistance, these compounds are not toxic to non-cancerous cells (Follit et al., 2015; Robey et al., 2008). The focus of this work is to find more compounds with these properties: efficient in neutralizing P-gp without causing harm to non-cancerous cells.

The datasets studied are provided by experts at CD4. They contain the characteristics of many different chemical compounds. The objective is to use these datasets to design models capable of reading similar data and predicting the efficiency

of compounds. Various data analytics methods are applied, on the one hand with the objective of predicting the compounds' exact level of P-gp inhibition (a continuous value) and on the other hand with the objective of predicting whether the compounds are effective in blocking P-gp or not (a binary value). In both cases, the models do not provide confidence in their ability to predict the target variable accurately. The models are nevertheless applied in order to obtain a list of the most promising compounds, which can be transmitted to the experts and studied further.

The primary objectives of this work are to obtain a list of potentially promising compounds and to answer the following research question: in the context of multidrug resistance caused by drug efflux, what are key characteristics of promising compounds?

In Chapter 2, the background knowledge that is necessary to understand this work is explained. In Chapter 3, the data provided by the experts at CD4 is presented and in Chapter 4, the methodology is detailed. Finally, the results are described in Chapter 5 and the conclusions and future work are developed in Chapter 6.

Chapter 2: Background and literature review

Cancerous cells have the ability to develop resistance (a lack of response) to traditional therapies. Although they are often initially sensitive to chemotherapy, they can develop this resistance over time through several mechanisms. One of these mechanisms is drug efflux. The adenosine triphosphate (ATP) is an organic compound that contains a large amount of chemical energy. Drug efflux uses ATP-binding cassette (ABC) transporters, proteins that transport a variety of substances across cellular membranes. In normal cells, this mechanism is beneficial: it keeps intracellular drug concentration below a cell-killing threshold. However, three transporters in particular protect cancerous cells from chemotherapy drugs by carrying them out of the cells: the Multidrug Resistance Protein 1 (MDR1), which produces P-gp, the Multidrug Resistance-Associated Protein 1 (MRP1) and the Breast Cancer Resistance Protein (BCRP).

2.1. Combination drug therapy

Using inhibitory drugs to block these proteins' activity can help sensitize cancerous cells to anticancer drugs. This concept is called combination drug therapy (Housman et al., 2014; Luqmani, 2005). Experiments have shown that inhibiting P-gp can reverse the effects of multidrug resistance by re-sensitizing cancerous cells to chemotherapeutics (Brewer, 2014; Follit et al., 2015). With this work, the aim is to help the discovery of such inhibitors by focusing on blocking P-gp’s activity in order to allow the drugs to remain in the target cells longer.

The objective of drug treatment is to destroy all the cancerous cells while inflicting minimum possible damage to the normal cells (Luqmani, 2005). This involves finding molecules that have the right characteristics to make acceptable drugs.

Combination drug therapy involves looking for candidate molecules that can block or activate a target protein, such as P-gp. The promising compounds are those that show binding activity towards this target protein.

2.2. The mechanisms of drug efflux

The mechanisms of drug efflux are illustrated in Figure 1. Transporters are efflux pumps situated on a cell's membrane. A binding domain is an area where compounds can bind to a protein. Two kinds of binding domains can be found on P-gp: the Nucleotide Binding Domains (NBDs) and the Drug Binding Domains (DBDs), sometimes also referred to as the Transmembrane Domains (TMDs). The two NBDs consume energy by absorbing the energy storage molecule ATP. This energy is used to power the two DBDs, which pump drugs out of the cell. A chemical compound that binds to an NBD hinders its activity by blocking the access for the energy molecules, thus preventing P-gp from consuming energy. However, a compound that binds to a DBD may be transported out of the cell by P-gp before it can have any effect.

In order to prevent drug efflux, an ideal compound would be one that strongly binds to the Nucleotide Binding Domains but not well to the Drug Binding Domains, because it would inhibit P-gp transport without being transported out of the cell.

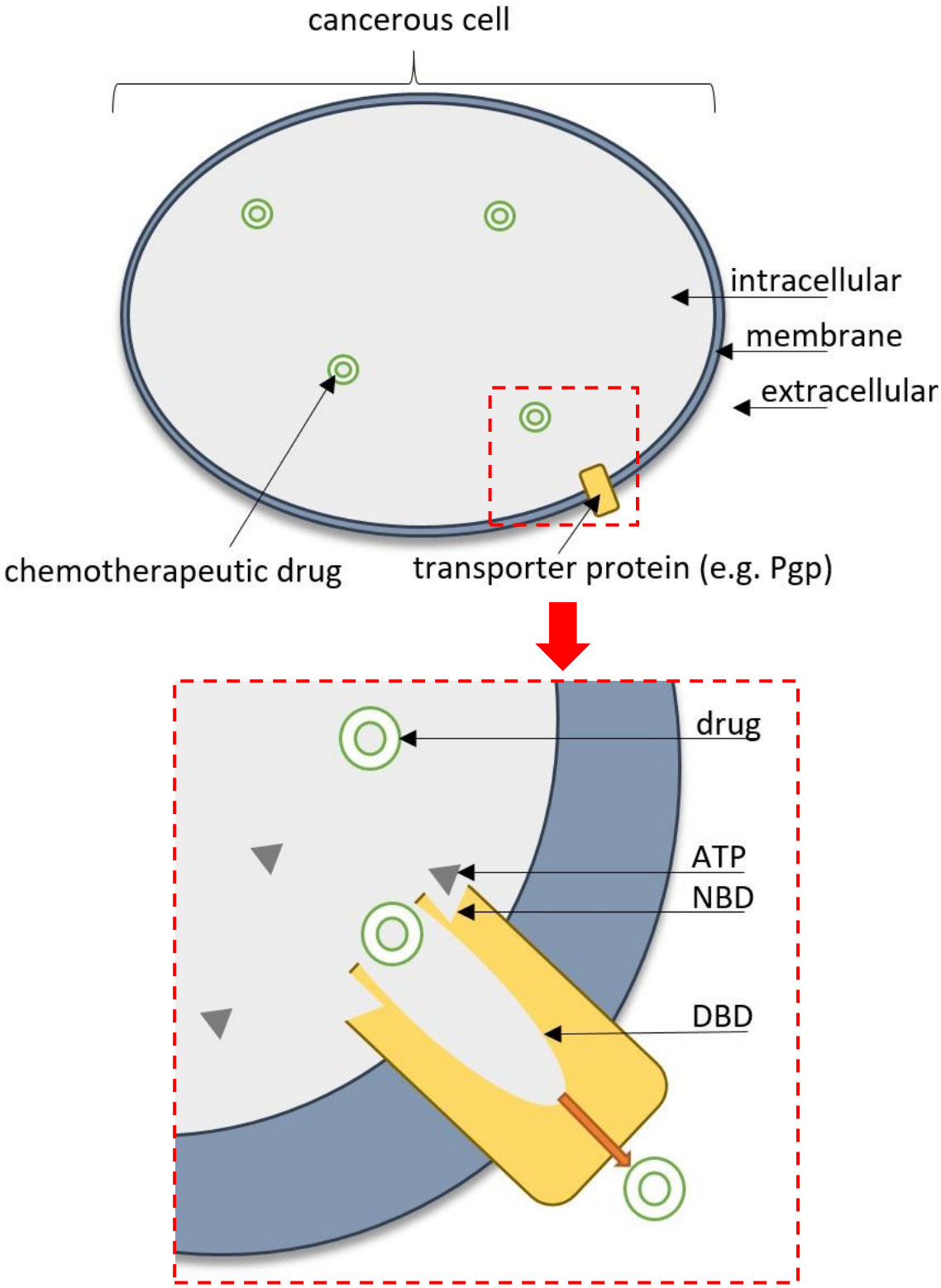

Figure 1 – Simplified schemas of the mechanisms of drug efflux

2.3. The P-glycoprotein

The transporter P-gp is particular because it is remarkably non-specific: it can transport a broad range of substrates across the plasma membrane, which is why an overexpression of P-gp causes multidrug resistance. The reason for this polyspecificity is unknown (Dolghih, Bryant, Renslo, & Jacobson, 2011). In humans, P-gp is expressed from the multidrug resistance gene, MDR1. The structural model of human P-gp that is used for this work was obtained by the researchers at CD4 by performing molecular dynamics experiments (Brewer, 2014).

P-gp is composed of two relatively symmetrical halves. Each half is composed of one DBD and one NBD. Each NBD is composed of six transmembrane helices. P-gp is illustrated in Figure 2 (original picture from *drugdiscoveryopinion.com*). The small area between the NBDs and the DBDs is referred to as the "Cyt" domain.

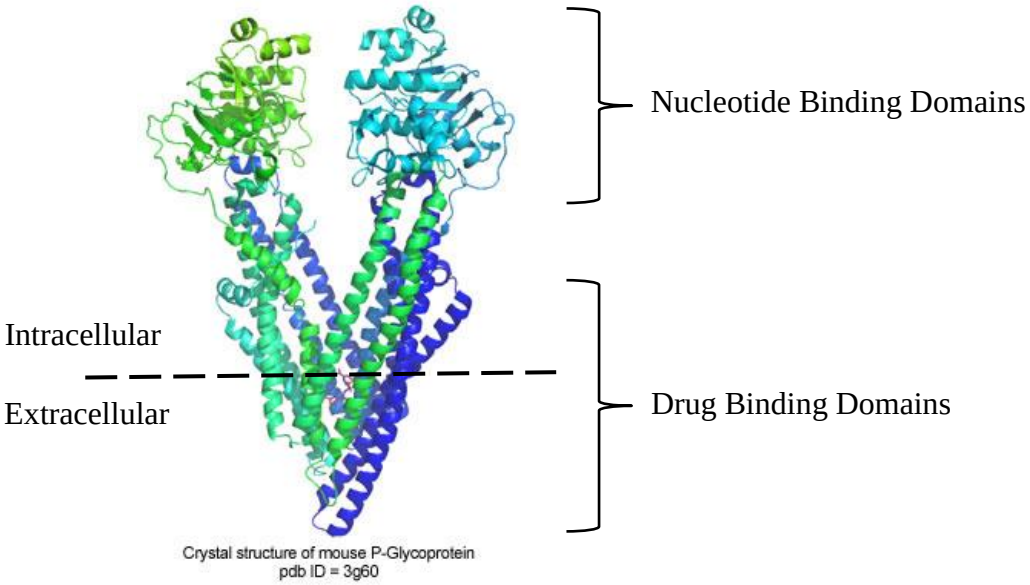

Figure 2 – The P-glycoprotein

The mechanism of drug efflux detailed previously can be observed in the specific case of P-gp in Figure 3 (Li et al., 2016). In the initial position of the protein, the two NBDs (located in the bottom part of the protein in Figure 3) are open to the energy molecules ATP and the DBDs (located in the top part of the protein and called TMDs in Figure 3) are open to the inside of the cell and ready for a substrate to interact with P-gp. When this happens, both the NBDs and the DBDs close in order to hold the ATP and the substrate inside the protein. This change of position causes P-gp to twist and rotate, until the DBDs open to the extracellular space. This new position allows the substrate to be pushed out of the cell and released by P-gp. Finally, the ATP molecules are hydrolyzed, the energy is consumed and P-gp goes back to its initial position. This process is called P-gp's catalytic transition cycle, and each position taken by the protein is called a pose (McCormick, Vogel, & Wise, 2015).

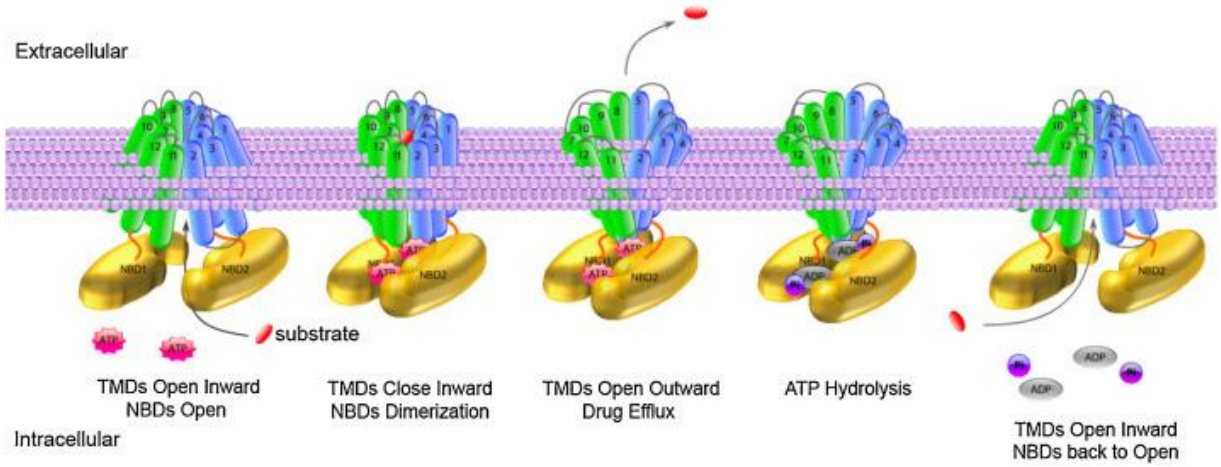

Figure 3 – The P-glycoprotein's catalytic cycle

2.4. The screening of compounds

The most effective way to find promising compounds is to test chemical compounds by high-throughput screening, or HTS (Chapron & Larin, 2004-2016; *High Throughput Screening: Methods and Protocols*, 2016). As stated in (Hughes, Rees, Kalindjian, & Philpott, 2011), the HTS method consists in the screening of an entire compound library (a database of commercially available compounds) against a target. This method uses complex laboratory automation to realize *in vitro* tests of potential inhibitor molecules. There is no need to have prior knowledge of the nature of the molecules likely to have activity at the target. According to (Chapron & Larin, 2004-2016), the HTS method allows scientists to test hundreds of thousands of compounds a day. This method was used by CD4 to produce the datasets studied in this work.

However, performing HTS is expensive. As stated in (Chapron & Larin, 2004-2016), the solution to this problem is to perform "virtual screening" before performing HTS, which is a "real screening" method. Virtual screening is cheaper than real screening, and allows compounds that have not been purchased or synthesized yet to be tested *in silico*. It cannot replace real screening, but it can reduce the number of compounds to be tested by real screening. This is the method that has been chosen by the experts at the origin of this work: a list of potentially promising compounds is obtained by virtual screening, and these compounds are purchased in order to perform actual *in vitro* screening by HTS and find actual promising compounds. The compounds that are selected *in vitro* are then tested *in vivo* by the biologists during later processes of drug development, in order to evaluate their real-world ability to inhibit P-gp.

2.5. Data analytics in anticancer drug discovery

The field of big data analytics has started to play an important role in healthcare because it provides tools to manage and analyze the large amounts of data generated by the healthcare industry. The digital datasets are typically too large to be stored with traditional hardware and too complex to be easily managed with traditional software (Belle et al., 2015; Raghupathi & Raghupathi, 2014).

More specifically, data analytics methods are being used to support the fight against cancer. Machine learning methods such as neural networks, decision trees or support vector machines were applied to imaging data from radiation oncology (the use of high-energy irradiation to kill cancerous cells), such as mammogram images, with the objective of detecting anomalies (El Naqa, 2016). Machine learning methods were also used on patient data from different institutions with the objective of predicting the risk of cancer (Ow & Kuznetsov, 2016) or the outcome of radiotherapy in the treatment of prostate cancer (Coates, Souhami, & El Naqa, 2016).

There have not been attempts to use data analytics methods to predict whether a compound has the ability to block P-gp or not. However, there have been attempts to predict whether a compound is transported by P-gp (a substrate) or not (a non-substrate) by studying the structure of the compound and applying data analytics methods. It is a difficult task: even if the binding affinity is high, the compound may not be well transported by P-gp because the ratio $\frac{rate\ of\ transportation\ by\ P-gp}{rate\ of\ reuptake\ into\ cells}$ is unfavorable. In some cases, this makes it complex to confidently classify compounds as substrates or non-substrates (Bikadi et al., 2011).

Two of the classical issues in data analytics for healthcare are the “curse of high dimensionality” (a high number of variables) and the “number of variables >> number of samples” problem (El Naqa, 2016). These issues are encountered in this work as well, principally because one of the provided datasets has many more variables than samples.

COMPOUND	*p450_3a4_uneq*	*2hyd_nbd_dbd*	*transition_nbd*	*3b5z_nbd_1*	*3b5x_dbd*	*4ksb_dbd*
ZINC00000017	1.23E-05	1.23E-05	8.80E-06	7.43E-06	1.23E-05	5.30E-06
ZINC00000190	8.80E-06	5.30E-06	6.28E-06		3.78E-06	7.43E-06
ZINC00000357	2.42E-05	2.04E-05	1.38E-06		5.63E-05	1.73E-05
ZINC00000534	1.04E-05	8.80E-06	3.20E-06			3.78E-06
ZINC00000757	0.0001105	4.75E-05	5.30E-06	6.66E-05	6.66E-05	7.89E-05
ZINC00000842	8.80E-06	1.23E-05	3.20E-06		8.80E-06	3.20E-06
ZINC00000868	4.75E-05	2.42E-05	3.39E-05	4.02E-05	0.0001105	0.0001105
ZINC00000982	7.89E-05	4.75E-05	2.04E-05	3.39E-05	7.89E-05	6.66E-05

Table 1 – Excerpt: virtual screening dataset

COMPOUND	SMILES	Rotatable Bonds	transition_nbd	xlogP	In Sicilo P-gp	P450	BCRP	Pgp-15	Toxicity
ZINC12274455	Cc1c(nnn1c2ccc3c(c2)c(on3	4	3.166234e-09	4.43	TRUE	-0.3	3.07398629	0.45594521	0.13429541
ZINC02312120	Cc1c2c(n(n1)C(=O)c3ccc4c(c	1	5.252385e-09	3.85	FALSE	0.15	1.96107819	-0.01697078	-0.06336148
ZINC12389647	c1ccc2c(c1)c(ns2)N3CCN(CC	3	1.220991e-08	3.83	TRUE	0.3	4.41645173	0.38448551	0.00778294
ZINC12274469	Cc1c(nnn1c2ccc3c(c2)c(on3	6	2.025471e-08	4.42	TRUE	0.2	3.955	0.6178538	0.06980065
ZINC12697057	Cc1cc(=O)c(nn1c2ccccc2F)C	3	1.150578e-09	4.26	FALSE	0.12	3.93412339	0.24765073	0.04670127
ZINC09349235	Cc1ccc(cc1)S(=O)(=O)c2ccc(	4	5.252385e-09	2.21	FALSE	0.45	3.19072848	0.02297204	-0.0123467
ZINC32848234	c1ccc2c(c1)c(n[nH]c2=O)CC(	6	1.220991e-08	4.01	FALSE	0.15	3.0688	0.54066271	0.26743845
ZINC49004308	c1cc(ccc1c2cc3c(=O)n(ncn3	4	8.713048e-09	3.88	FALSE	0.1	1.1959	0.02365146	0.00158615

Table 2 – Excerpt: real screening dataset

Chapter 3: Data

3.1. "Combination drug therapy" data

Two datasets are provided by the experts at CD4.

3.1.1. Results of the virtual screening

The first dataset contains a list of 159,000 chemical compounds and the values they take for 15 features. The values were obtained by virtual screening. By analyzing the results, the experts at CD4 selected 31 compounds that they believed to be promising. These 31 compounds were purchased in order to perform real screening (by HTS) and obtain accurate values. As has been explained previously (see Chapter 2: Section 2.4), performing virtual screening first allowed the experts to reduce the number of compounds to purchase and analyze by HTS.

Table 1 is an excerpt of this first dataset. The complete list of variables in this dataset is available in Appendix A. All 15 variables are numerical. They contain information on the strength of the binding activity between the compounds and various receptors: a low value indicates a strong binding between the compound and the receptor, whereas a high value indicates a weak binding.

3.1.2. Results of the real screening

The second dataset contains the values obtained by HTS for 77 compounds: the 31 compounds from the virtual screening dataset as well as 46 compounds selected by CD4.

Table 2 is an excerpt of this second dataset. The complete list of variables in this dataset is available in Appendix B. HTS provided information on 89 features of the 77 compounds. Among these features, 5 are integer, 65 are numerical (they contain

decimal numbers), 11 are factors (they contain characters), 1 is logical (it contains a binary value) and 7 are empty. These variables are more diversified than the variables from the virtual screening dataset: some of them contain information on the strength of the binding activity between the compounds and various receptors, like in the virtual screening dataset, and others describe the structure of the compound. In the case of the variables containing information on binding activity, a low value indicates a strong binding between the compound and the receptor, whereas a high value indicates a weak binding.

Four of the variables in this dataset are target features (the 4 variables on the right in Table 2): an ideal compound takes a high value for some and a low value for others. The target features are:

- **P-gp 15 micromolar efficacy:** represents the ability of the chemotherapeutic used with the compound to kill cells that are resistant because of P-gp, normalized to the ability of the chemotherapeutic used without the compound. This numerical variable takes values from 0 to 1 (although a few values fall slightly below 0, which may be due to the equipment used to test the compounds). A value of 0 means no difference with or without the compound. A **high** value indicates a promising compound;
- **10 micromolar BCRP fold inhibition:** represents the ability of the chemotherapeutic used with the compound to kill cells that are resistant because of BCRP. This numerical variable takes values from 1 to 5. A **high** value indicates a promising compound;

- **10 micromolar P450 inhibiton:** represents the inhibition of the compound against the target P450. P450_3A4 is an important enzyme that should not be disrupted to avoid side effects. This numerical variable takes values from 0 to 1. It is better for the patient if this value is **close to 0**;
- **Actual toxicity:** represents the ability of the compound by itself to kill cells. This numerical variable takes values from 0 to 1. This value must be **low**: if it is higher than 0.5, the compound cannot be accepted because the compound's action is through killing the cell itself, not sensitizing it so that the chemotherapeutic can kill it. Ideally, the compound should have no effect without the chemotherapeutic.
- **Multifactor usability**: an indicator variable provided by the experts on the compound's overall effectiveness. It is based on the other 4 targets. This variable takes the value "True" if the compound turned out to be promising during the *in vitro* tests and the value "False" otherwise.

According to CD4, either "P-gp 15 micromolar efficacy", "Multifactor usability" or "10 micromolar BCRP fold inhibition" having a satisfactory value is enough for a compound to be promising. While they are important characteristics of a compound too, "Actual toxicity" and "10 micromolar P450 inhibition" can be improved during later processes of drug development and optimization.

The subject of this work is the transporter P-gp, which is why the target variables "P-gp 15 micromolar efficacy" and "Multifactor usability" are the focus. As can be observed in Figure 4, which illustrates the distribution of "P-gp 15 micromolar

efficacy", most compounds are not extremely efficient in blocking P-gp, even though they were selected after the virtual screening step because they were believed to be promising. This deduction is confirmed in Figure 5, which illustrates the values taken by the target variable "Multifactor usability": most of the potentially promising compounds are not actually efficient in blocking P-gp.

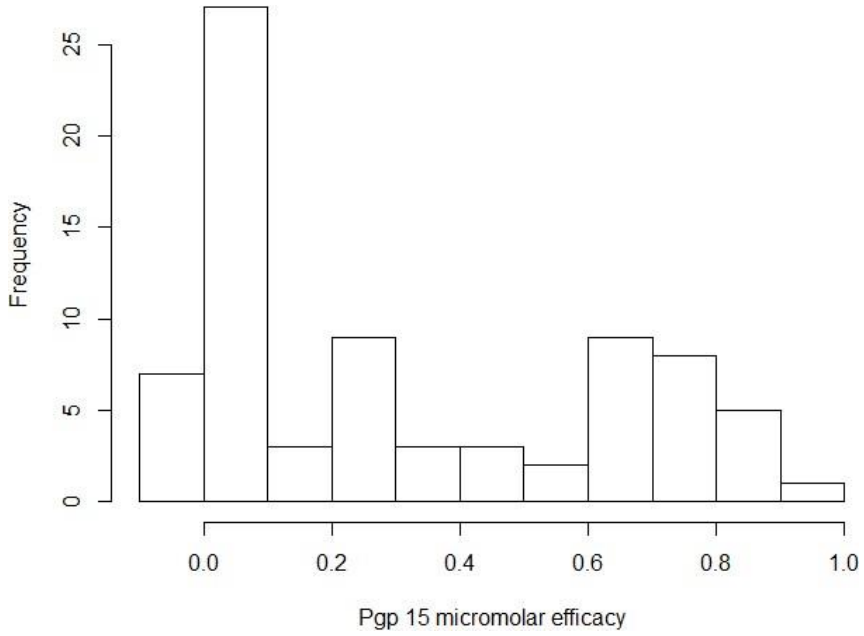

Figure 4 – Distribution of the target variable "P-gp 15 micromolar efficacy"

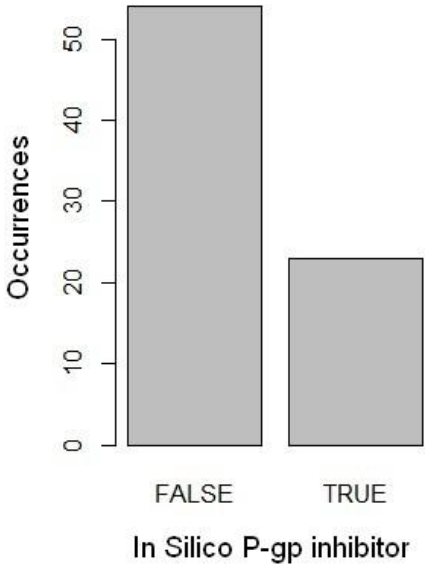

Figure 5 – Distribution of the target variable "Multifactor usability"

3.2. “AutoDock Vina” data

The third source of available data contains the results of experiments on P-gp docking. This data was obtained by CD4 using the open-source software AutoDock Vina, developed by Dr. Oleg Trott at the Script Research Institute of San Diego, California.

3.2.1. Molecular docking

Molecular docking is a computational method of prediction of the preferred orientation of a molecule when it binds to another molecule, as well as the binding affinity. Knowing the preferred orientation of a molecule can be useful to predict the strength of the binding between two molecules.

The objective of AutoDock Vina is to perform molecular docking. This software is especially effective for protein-ligand docking: the prediction of the position and orientation of a small molecule when it binds to a protein or enzyme. This method is particularly useful in the context of drug discovery: it is used to screen virtual libraries of molecules in order to obtain clues about which molecules are promising for further drug development. It allows researchers to determine a compound’s orientation when it binds to a target protein, and to calculate its affinity and level of activity with this target protein (Kitchen, Decornez, Furr, & Bajorath, 2004; Trott & Olson, 2010).

The researchers at CD4 used AutoDock Vina to virtually dock a library of compounds against P-gp. 123 compounds were tested against 8 structures that represent the stages of P-gp’s catalytic transition cycle, which was detailed previously in Chapter 2 and illustrated in Figure 3. These structures are called poses of the protein. Figure 6 (provided by CD4) depicts five of P-gp’s poses, as well as their names.

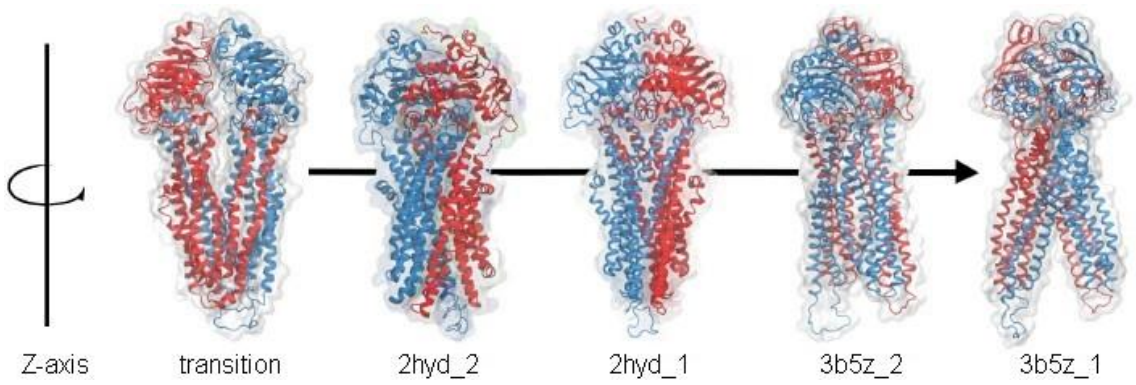

Figure 6 – Stages of P-gp's catalytic cycle

The compounds were actually tested against 28 receptors:

- 3 receptors for each of the 8 poses of P-gp: one receptor on the Drug Binding Domain, one on the Nucleotide Binding Domain and one in the "Cyt" domain, which covers the region between the DBD and the NBD;
- 3 receptors (DBD, NBD and Cyt) on MDR1_6, the Multi-Drug Resistance Protein 1, which is similar to the 2_hyd structures of P-gp;
- 1 receptor on BCRP, the Breast Cancer Resistance Protein.

3.2.2. Types of data

For each of the 3,444 compound-receptor couples, the result of the AutoDock Vina analysis is the top 20 docking positions, numbered 1 to 20 with the position number 1 being the best docking position on the protein. For each docking position, three types of data are available:

- **The estimated Kd:** the binding affinity.

This value represents how much of a molecule is needed to have a potential inhibitory effect. For example, if the compound X has an estimated Kd of 2.025 x 10E-8 when tested against the receptor Y_NBD, then a concentration of 2.025 x 10E-8 Mol is needed for the compound to have an inhibitory effect. In drug design, it is optimal for the Kd to be as low as possible, so that it will not be necessary to use a large quantity of the drug for it to take effect.

Since we want the compounds to bind strongly to the NBDs and weakly to the DBDs (see Chapter 2: Section 2.2), we want to observe a low Kd for the NBD receptors and a high Kd for the DBD receptors. For example, if the compound X mentioned previously has an estimated Kd of 4.708 x 10E-8 when tested against the receptor Y_DBD, then the ratio $\frac{Kd\ for\ DBD}{Kd\ for\ NBD}$ is 2.32. We want that ratio to be as high as possible so that the molecules interacts preferentially with the NBD over the DBD.

- **The estimated deltaG (dG):** the change in Gibbs free energy.

The dG is a quantitative measure of a reaction's favorability at constant temperature and pressure. It is measured as a change in Gibbs free energy. A low dG means that the reaction is favorable. We are looking for molecules whose interaction with the NBD is more favorable than their interaction with the DBD: just like in the case of the Kd, we want the ratio $\frac{dG\ for\ DBD}{dG\ for\ NBD}$ to be as high as possible.

The dG is the value that is used to evaluate the binding strength of each docking position, in order to select the top 20 docking positions (those with the lowest dG) for each compound-receptor couple.

- **The coordinates of each atom that make up the molecule:**

From these coordinates, the center of mass (COM) coordinates can be calculated. This information is useful to determine if the molecule docks to the same spot of P-gp repeatedly or if the docking spot is highly variable. This can in turn be used to find out which spots on P-gp are creating favorable biochemical interactions.

My hypothesis is that a compound that docks repeatedly to the same spot on the Drug Binding Domain of P-gp may be more easily transported out of the cell, whereas a compound that docks repeatedly to the same spot on the Nucleotide Binding Domain of P-gp may be more efficient in reversing the multidrug resistance.

```
MODEL 1
REMARK VINA RESULT:     -10.1      0.000      0.000
REMARK  6 active torsions:
REMARK  status: ('A' for Active; 'I' for Inactive)
REMARK    1  A    between atoms: C7_7  and  O3_11
REMARK    2  A    between atoms: O3_11  and  C9_12
REMARK    3  A    between atoms: C9_12  and  C10_14
REMARK    4  A    between atoms: C14_19  and  N2_20
REMARK    5  A    between atoms: N2_20  and  C15_21
REMARK    6  A    between atoms: C19_25  and  C21_27
ROOT
ATOM      1  C10 <0> A   1       1.399  -3.320  31.093  1.00  0.00     0.085 A
ATOM      2  C11 <0> A   1       0.858  -4.252  30.198  1.00  0.00     0.023 A
ATOM      3  C12 <0> A   1      -0.089  -5.144  30.672  1.00  0.00     0.020 A
ATOM      4  C13 <0> A   1      -0.472  -5.090  32.001  1.00  0.00     0.119 A
ATOM      5  N1  <0> A   1       0.050  -4.204  32.827  1.00  0.00    -0.242 NA
ATOM      6  C14 <0> A   1       0.963  -3.332  32.429  1.00  0.00     0.144 A
ENDROOT
BRANCH   1   7
ATOM      7  C9  <0> A   1       2.409  -2.349  30.641  1.00  0.00     0.299 C
ATOM      8  O4  <0> A   1       2.234  -1.160  30.826  1.00  0.00    -0.259 OA
BRANCH   7   9
ATOM      9  O3  <0> A   1       3.528  -2.776  30.023  1.00  0.00    -0.281 OA
BRANCH   9  10
ATOM     10  C7  <0> A   1       4.224  -1.828  29.212  1.00  0.00     0.315 A
ATOM     11  C5  <0> A   1       4.858  -2.525  28.034  1.00  0.00     0.013 A
ATOM     12  C4  <0> A   1       6.234  -2.305  28.148  1.00  0.00     0.050 A
ATOM     13  C6  <0> A   1       4.357  -3.259  26.985  1.00  0.00     0.013 A
ATOM     14  C3  <0> A   1       7.095  -2.837  27.186  1.00  0.00     0.018 A
ATOM     15  C8  <0> A   1       6.458  -1.496  29.358  1.00  0.00     0.296 A
ATOM     16  C2  <0> A   1       6.581  -3.572  26.138  1.00  0.00     0.001 A
ATOM     17  C1  <0> A   1       5.218  -3.783  26.036  1.00  0.00     0.001 A
ATOM     18  O1  <0> A   1       5.287  -1.231  29.959  1.00  0.00    -0.281 OA
ATOM     19  O2  <0> A   1       7.545  -1.123  29.755  1.00  0.00    -0.259 OA
ENDBRANCH   9  10
ENDBRANCH   7   9
ENDBRANCH   1   7
BRANCH   6  20
ATOM     20  N2  <0> A   1       1.487  -2.425  33.332  1.00  0.00    -0.340 N
ATOM     21  H9  <0> A   1       2.408  -2.133  33.247  1.00  0.00     0.167 HD
BRANCH  20  22
ATOM     22  C15 <0> A   1       0.689  -1.928  34.366  1.00  0.00     0.034 A
ATOM     23  C20 <0> A   1      -0.533  -1.333  34.077  1.00  0.00     0.045 A
ATOM     24  C19 <0> A   1      -1.322  -0.847  35.102  1.00  0.00     0.050 A
ATOM     25  C18 <0> A   1      -0.891  -0.942  36.412  1.00  0.00     0.016 A
ATOM     26  C17 <0> A   1       0.327  -1.528  36.703  1.00  0.00     0.003 A
ATOM     27  C16 <0> A   1       1.119  -2.021  35.683  1.00  0.00     0.029 A
BRANCH  24  28
ATOM     28  C21 <0> A   1      -2.651  -0.207  34.791  1.00  0.00     0.417 C
ATOM     29  F1  <0> A   1      -2.765   0.996  35.494  1.00  0.00    -0.166 F
ATOM     30  F2  <0> A   1      -2.737   0.045  33.419  1.00  0.00    -0.166 F
ATOM     31  F3  <0> A   1      -3.684  -1.071  35.173  1.00  0.00    -0.166 F
ENDBRANCH  24  28
ENDBRANCH  20  22
ENDBRANCH   6  20
TORSDOF 6
ENDMDL
```

Figure 7 – Excerpt: file describing the structure of the compound ZINC00601275_SMU113 when docked to the receptor 2hyd_1_cyt

3.3. Data cleaning

The data is not usable as it is:

- The Combination drug therapy data contains several non-numeric variables that are not useful to build models, such as the variable “Smiles”: a representation in ASCII characters of the chemical structure of the compounds (see Table 2). It also contains many missing values. Additionally, the dataset containing the results of the real screening is unbalanced: it does not contain enough samples for the number of variables (only 77 samples for 89 variables).
- The AutoDock Vina data is scattered in many different files: one folder for each receptor, containing:
  - One file for each docked compound describing the structure and the coordinates of each atom. An excerpt of such a file is presented in Figure 7;
  - One file for each docked compound describing each of the top 20 docking positions. This file is structurally similar to the previous one;
  - One file containing the Kd and dG for each compound-receptor couple. An excerpt of such a file is presented in Figure 8. The names of the compound and receptor can be read in the first column, the dG in the second column and the Kd in the third column.

This is why a preliminary step of preprocessing data is necessary. The first treatment of the data consists in removing empty and non-numerical variables from the Combination drug therapy data.

```
./ZINC08685602_SMU30_2hyd_dbd_1_ori_out.pdbqt -8.6 .00000049966444350388
./ZINC10363190_SMU35_2hyd_dbd_1_ori_out.pdbqt -7.9 .00000162770880154540
./ZINC07069240_SMU115_2hyd_dbd_1_ori_out.pdbqt -8.0 .00000137500896892478
./ZINC09835648_SMU21_2hyd_dbd_1_ori_out.pdbqt -9.2 .00000018157303824518
./ZINC48237631_SMU88_2hyd_dbd_1_ori_out.pdbqt -8.2 .00000098121271932813
./ZINC24781309_SMU54_2hyd_dbd_1_ori_out.pdbqt -8.9 .00000030120689087418
./ZINC04741719_SMU16_2hyd_dbd_1_ori_out.pdbqt -8.1 .00000116154048121417
./ZINC25054259_SMU84_2hyd_dbd_1_ori_out.pdbqt -8.1 .00000116154048121417
./ZINC12446716_SMU23_2hyd_dbd_1_ori_out.pdbqt -8.2 .00000098121271932813
./ZINC06868070_SMU67_2hyd_dbd_1_ori_out.pdbqt -9.0 .00000025444488354472
```

Figure 8 – Excerpt: file describing the results (Kd and dG) of the docking of each compound to the receptor 2hyd_1_cyt

As can be observed in Table 3, this treatment significantly reduces the number of missing values as well as the number of variables, particularly in the case of the dataset containing the results of the real screening. Additionally, the proportion of missing values for each variable in the virtual screening dataset and in the real screening dataset after this first treatment can be visualized in Figure 9 and in Figure 10 respectively. A lot of missing values are still in the real screening dataset (more than 60% for more variables),

The second treatment of the data consists in removing constant and almost-empty variables, as well as the observations for which most values are missing. The variables that are present in the real screening dataset but absent from the virtual screening dataset are removed as well because they are useless for the prediction (this will be detailed in Chapter 4: Section 4.1 and Figure 12). The final real screening dataset contains 71 observations for 14 variables, including the 4 target variables, and less than 1% of missing values. Figure 11 shows the proportion of missing values for each variable of the real screening dataset after the second treatment.

Virtual screening dataset	*No treatment*	*Treatment 1*
Number of lines (observations)	158,867	158,867
Number of columns (variables)	17	13
Number of missing values	664,832	471,663
Percentage of missing values	24.62	22.84

Real screening dataset	*No treatment*	*Treatment 1*	*Treatment 2*
Number of lines (observations)	77	77	71
Number of columns (variables)	85	11	10
Number of missing values	3,739	354	7
Percentage of missing values	57.13	41.79	0.90

The number of columns and the percentage of missing values in the real screening dataset are calculated without taking into account the target variables, which contain no missing values.

Table 3 – Contents of the datasets

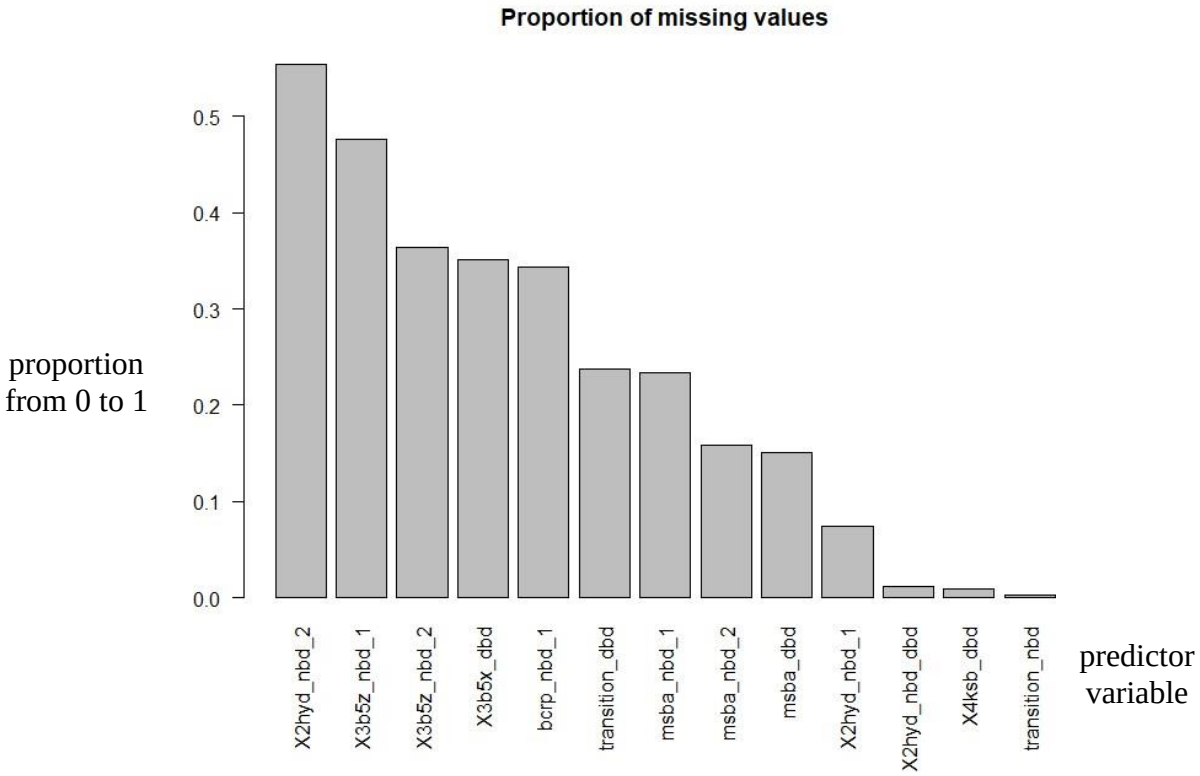

Figure 9 – Missing values in the virtual screening dataset (after treatment)

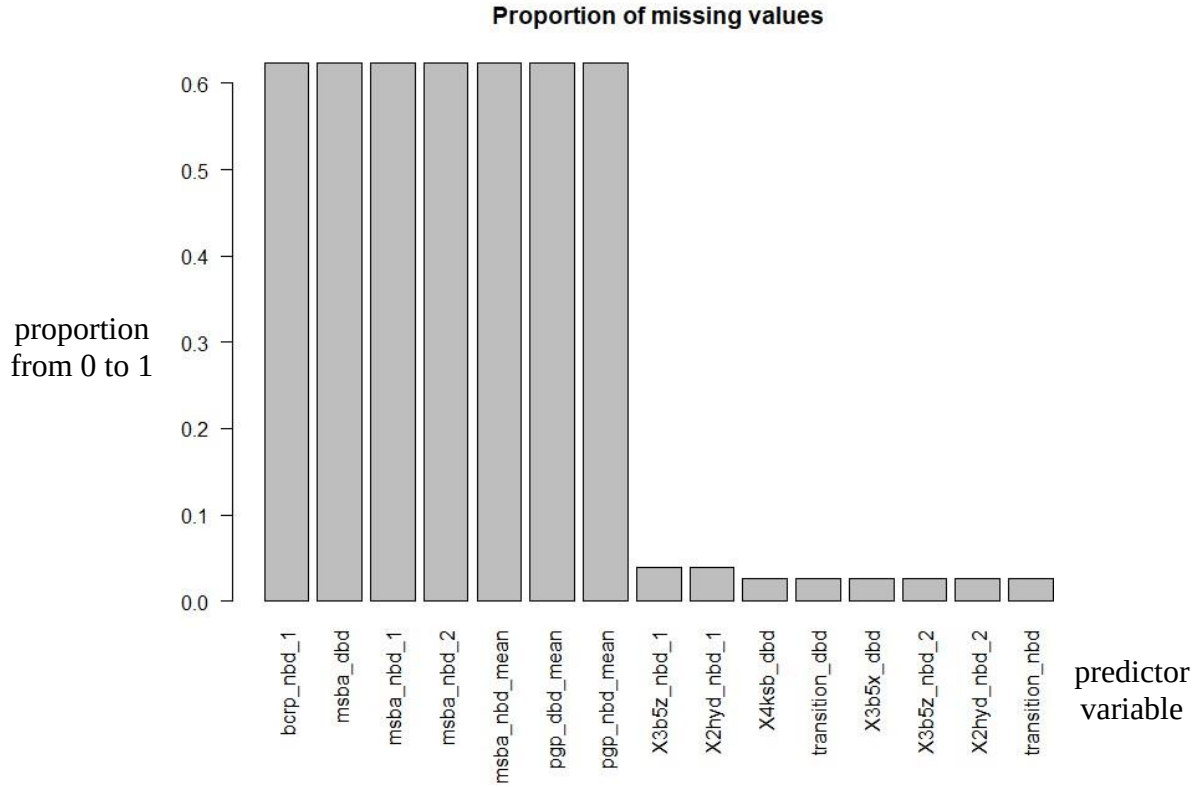

Figure 10 – Missing values in the real screening dataset (after first treatment)

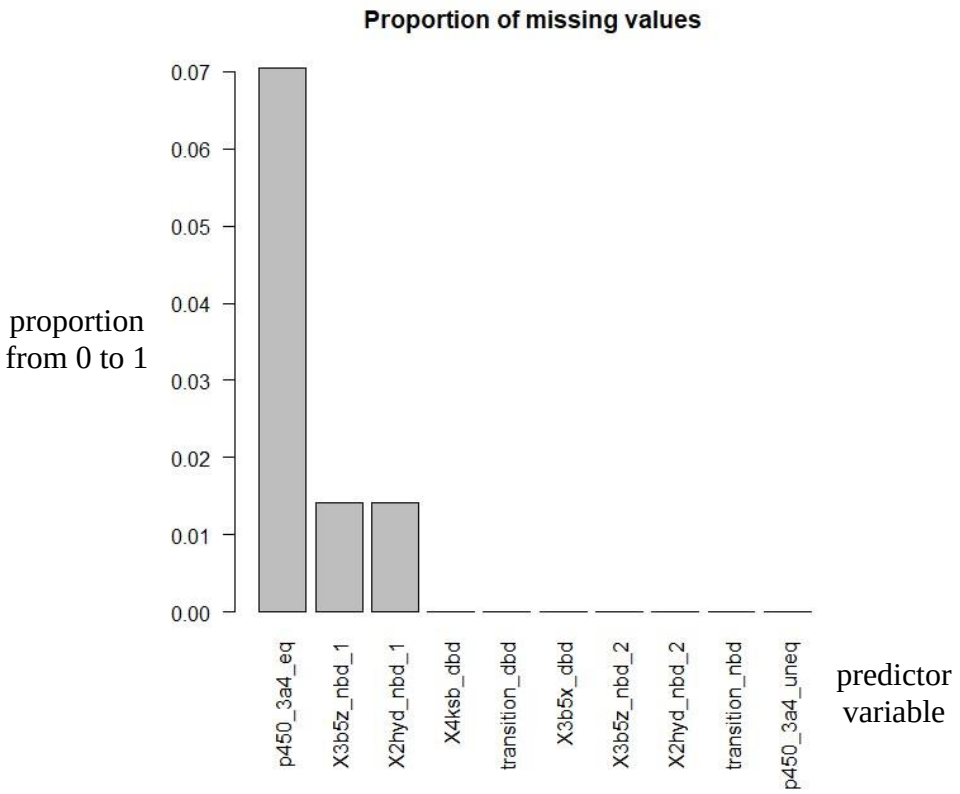

Figure 11 – Missing values in the real screening dataset (after second treatment)

The treatment of the AudoDock Vina data consists in creating a large dataset gathering the information previously scattered in many different files. The files containing information about the BCRP transporter or about the receptors in the "Cyt" domain of P-gp are discarded: only the information about "DBD" and "NBD" receptors on the glycoprotein and on MDR1_6 are processed. The final dataset contains 1,118,935 lines and 12 columns. Table 4 is an excerpt of this dataset.

COMPOUND	Receptor	dG	Kd	Dock_pose	Atom	x	y	z
ZINC00601275_SMU113	2hyd_dbd_1	-8.9	3.0120689087418e-07	1	C	-8.021	-15.289	6.449
ZINC00601275_SMU113	2hyd_dbd_1	-8.9	3.0120689087418e-07	1	C	-6.664	-14.940	6.452
ZINC00601275_SMU113	2hyd_dbd_1	-8.9	3.0120689087418e-07	...	...	...	-	...
ZINC00601275_SMU113	2hyd_dbd_1	-8.9	3.0120689087418e-07	20	C	-0.295	-5.392	3.469
ZINC14250428_SMU50	3b5x_dbd	-9.4	1.2957135454363e-07	1	C	6.565	-7.645	-10.244
ZINC14250428_SMU50	3b5x_dbd	-9.4	1.2957135454363e-07	1	O	6.468	-8.989	-10.417
ZINC14250428_SMU50	3b5x_dbd	-9.4	1.2957135454363e-07	...	...	...	...	...
ZINC14250428_SMU50	3b5x_dbd	-9.4	1.2957135454363e-07	20	N	4.423	-6.527	-10.585
...	...	...	...	...	...	...	...	...

Table 4 – Excerpt: dataset containing the AutoDock Vina results

Chapter 4: Methodology

4.1. Objective

The focus of this work is to find relationships between two of the main target features, "P-gp 15 micromolar efficacy" and "Multifactor usability", and the predictor features (the 85 features from the real screening dataset that are not target features). The objective, schematically represented in Figure 12, is:

- To use the real screening dataset (which is the training set) to design models capable of predicting the values taken by the target variables ("P-gp 15 micromolar efficacy" in the example of Figure 12). Only the variables common to both datasets can be used;
- To apply these models to the virtual screening dataset in order to predict the compounds' efficacy;
- To use the predicted values to identify new promising compounds, which can later be bought by the experts at CD4 and analyzed by real screening.

If the hypothesis stated in Chapter 3: Section 3.2.2 can be confirmed (a compound that docks repeatedly to the same spot on the DBD is more easily transported out of the cell, whereas a compound that docks repeatedly to the same spot on the NBD is more efficient in reversing the multidrug resistance), the information obtained from the AutoDock Vina dataset can be combined to the information obtained from the Combination drug therapy dataset to select the promising compounds more accurately.

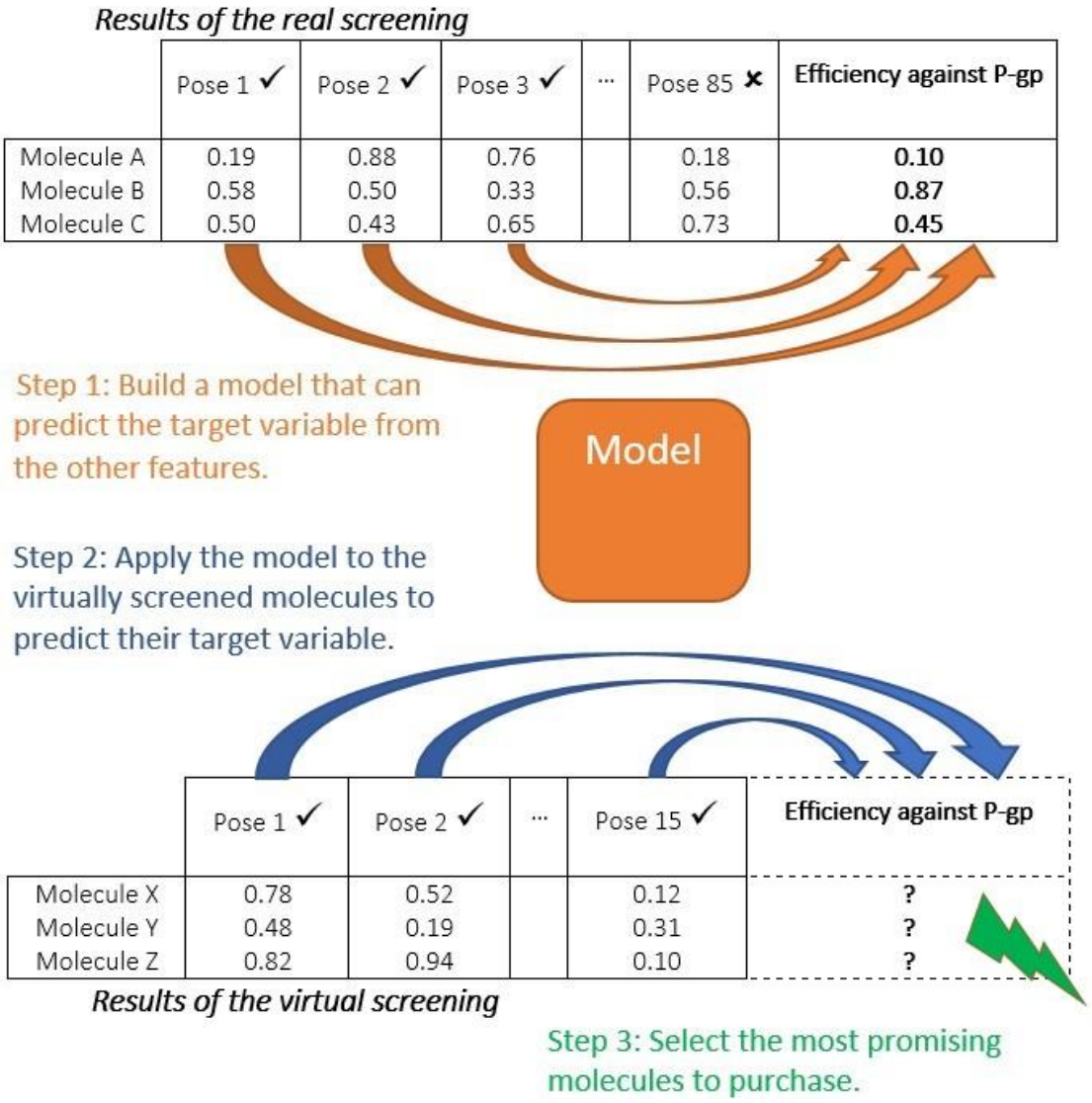

Figure 12 – Schematized objective

4.2. Tools

Various programming languages are well-adapted to data analysis. Among them are Python and R, two of the most used languages in this field. In this work, Python is used occasionally (for the creation of the AutoDock Vina dataset mentioned in Chapter 3: Section 3.3) but the main language is R, which is a well-documented language with many available libraries.

Most of the R libraries that are used in this work are model building libraries that implement data analytics methods. I use the method “lm” from the library “stats”, the method “stepAIC” from the library “MASS”, the method “rpart” from the library of the same name, the method “randomForest” from the library of the same name, the method “nnet” from the library “caret” and the method “svm” from the library “e1071”. The libraries “hydroGOF”, “rfUtilities” and “caTools” provide the quality assessment functions that are used to evaluate the models.

4.3. Machine learning

In order to identify some key characteristics of promising compounds, I apply regressions to the real screening dataset, as well as various machine learning methods: algorithms that learn from data without being explicitly programmed.

There are two main types of machine learning algorithms:

- The supervised learning algorithms, which produce models from a training dataset in which each observation is labeled: the values taken by the target variable are known and they guide the learning process. For example, a problem of image classification fall under the category of supervised learning problems;

- The unsupervised learning algorithms, whose objective is to describe unlabeled data. For example, cluster analysis algorithms such as k-means fall under the category of unsupervised learning problems.

The difference between these two types of problems is illustrated in Figure 13:

- In the case of supervised learning, the observations in the dataset are labeled "A" or "B" and the model learns to recognize these two categories depending on the values taken by the variables x1 and x2. When the model is applied to a new, unlabeled observation (labeled "X" in Figure 13), it is able to identify the most appropriate category by analyzing the values taken by x1 and x2;
- In the case of unsupervised learning, the observations are not labeled. By analyzing the values taken by x1 and x2, the model estimates the number of represented categories to two. It associates each category to a set of values for x1 and a set of values for x2. When the model is applied to a new, unlabeled observation, it is able to identify the most appropriate category by analyzing the values taken by x1 and x2.

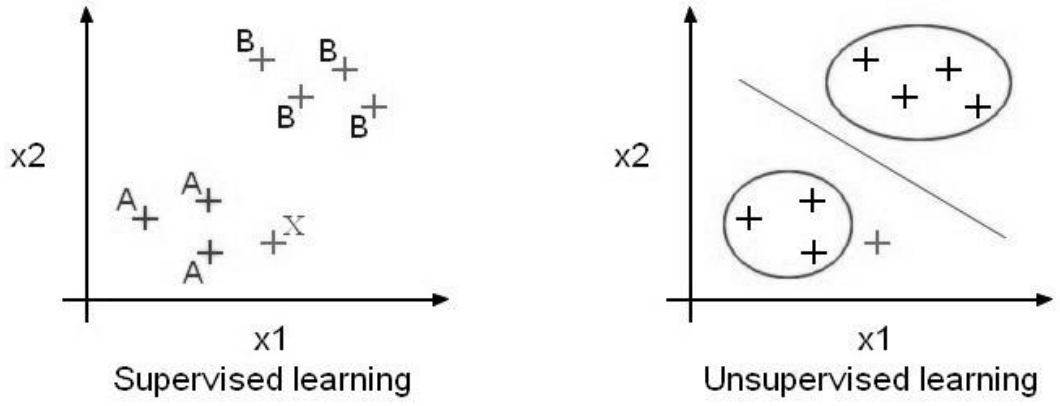

Figure 13 – Concept of supervised versus unsupervised learning

The problem dealt with in this work falls under the category of supervised learning. The observations are labeled: in the real screening dataset, the values taken by the target variables are known. The new, unlabeled observations that the model has to classify after is has been trained are the observations from the virtual screening dataset.

A common approach in the field of supervised learning is the partition of the training dataset into two sub-datasets: a learning dataset (which represents most of the time 70% of the original dataset) and a testing dataset (the remaining 30%). This allows the training of the model on a sufficient number of observations, before estimating its quality by applying it to the testing data and comparing the predicted values with the actual values taken by the target variable.

It is also possible to divide the dataset into three sub-datasets: learning (production of a model using various approaches), validation (selection of the best approach) and testing (measure of the quality of the selected approach) datasets.

In both cases, it is essential for all datasets to be balanced: each must be an accurate representation of the original dataset.

In the case of this work, the training dataset (the real screening dataset) only contains 71 observations after the data cleaning step. This number of observations, which is already too low to obtain good-quality models, becomes way too low if the dataset is partitioned. The tests made with a “learning dataset / testing dataset” partition are not conclusive because the models did not have access to enough data to learn properly.

For this reason, the Leave-One-Out Cross-Validation (LOOCV) method is used to evaluate the models. In this method, the partition of the training dataset into a learning dataset and a testing dataset is slightly different: only one observation is selected in the testing dataset. The LOOCV algorithm proceeds as follows:

1. For each observation *o* from the real screening dataset:
   - Remove *o* from the training dataset;
   - Train the model;
   - Predict the value taken by the target variable for *o*;
   - Assess the quality of the model by comparing the prediction to the actual value.
2. Assess the quality of the final model by computing the average quality of all the intermediate models;
3. Train the final model on the whole training dataset.

4.4. Methods

I used the programming language R and various data analytics methods to produce models capable of predicting the target variables. These data analytics methods are detailed below.

- The **linear regression**:

This approach models the link between a dependent variable Y and one or more explanatory variables X, using an equation of the form $Y = a * X + b$.

- The **stepwise regression**:

The stepwise regression is an improved linear regression: the algorithm includes an automatic process of explanatory variable selection. At a given step, the algorithm searches for the independent variable that optimizes a criteria given the variables already selected. In this work, the criteria used to select variables is the Akaike information criterion (AIC), an estimator of the quality of statistical models. Only the set of variables that were selected are used to build the model (Zhang, 2016).

- The **decision tree:**

This method uses a tree-like model to represent decisions and their consequences. Each internal node of the tree represents a test on an attribute of the observation (for example, “Is the value taken by the variable X lower than 1?”). Each branch represents one outcome of the test (for example, “Yes” or “No”). Each leaf represents the decision of associating a specific label to the observation.

- The **random forest**:

The random forest is a group of decision trees, whose output is the mode (for classification models) or the mean (for regression models) of the individual predictions made by each tree. In the case of the example random forest in Figure 14 (picture from unknown source), four sub-datasets were randomly selected in order to build four decision trees. In each tree, the nodes are built by randomly selecting a set of explanatory variables and by choosing the most relevant one. New observations are classified by vote: the category chosen by the majority of the individual trees is the category chosen by the forest (Liaw & Wiener, 2002).

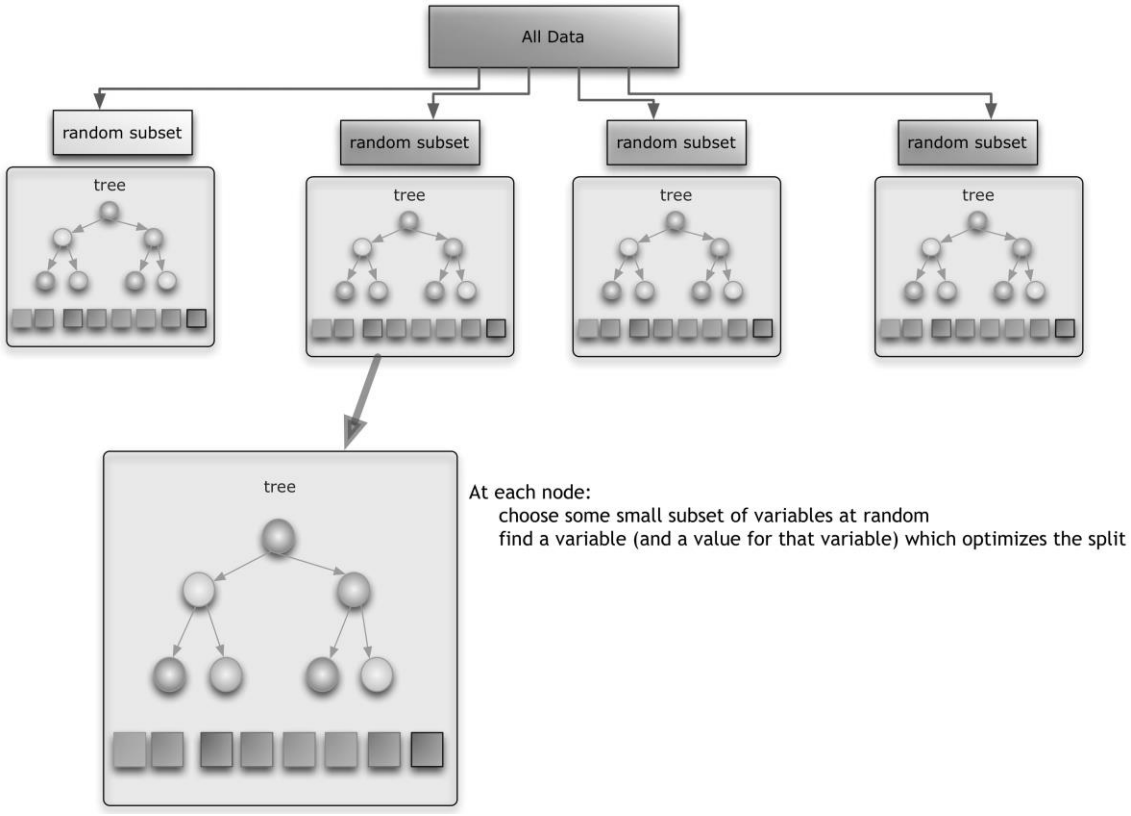

Figure 14 – Example of random forest

- The **neural network**:

This approach is inspired by the connections made in the animal brain and in the nervous system. It is a "black box" system (see Figure 15) whose inputs and outputs are known, but whose inner working are hidden.

Figure 15 – Concept of "black box" systems

As can be observed in the example in Figure 16 (picture from *supinfo.com*), a neural network takes one or more values as its input data, manipulates these values using the inner nodes contained in one or more hidden layers (which make up the black box whose inner workings are unknown) and returns one or more values as the output data.

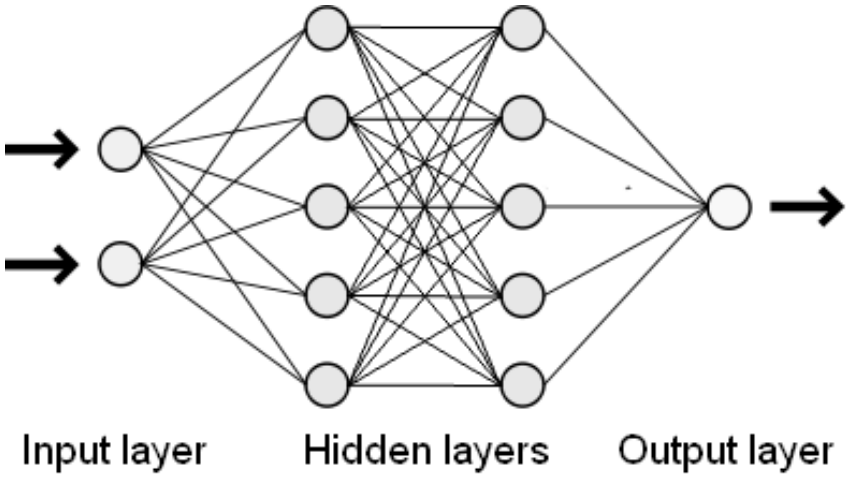

Figure 16 – Example of neural network

- The **support vector machine**:

This algorithm, which falls under the category of "black box" systems as well, classifies observations using a hyperplane. The concept is illustrated in Figure 17 and Figure 18 (pictures from *irisa.fr*) for a two-dimensional problem. The first step, illustrated in Figure 17, is to find a space in which the data points are linearly separable and to map the data points in that new space. The next step, illustrated in Figure 18, is to look for the "best hyperplane", which separates the categories of data properly while maximizing the margin (the distance from the closest data point to the hyperplane). The hyperplane can be determined using points called "support vectors".

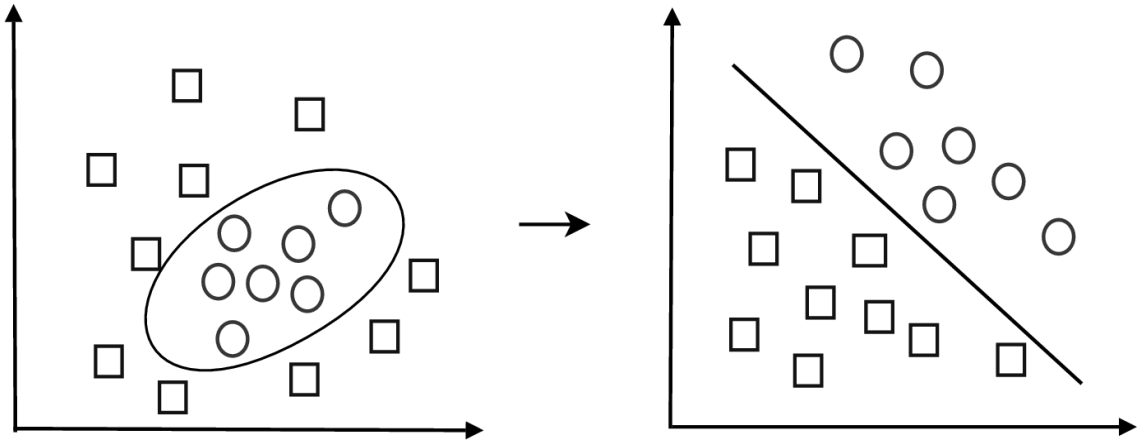

Figure 17 – SVM: projection of data

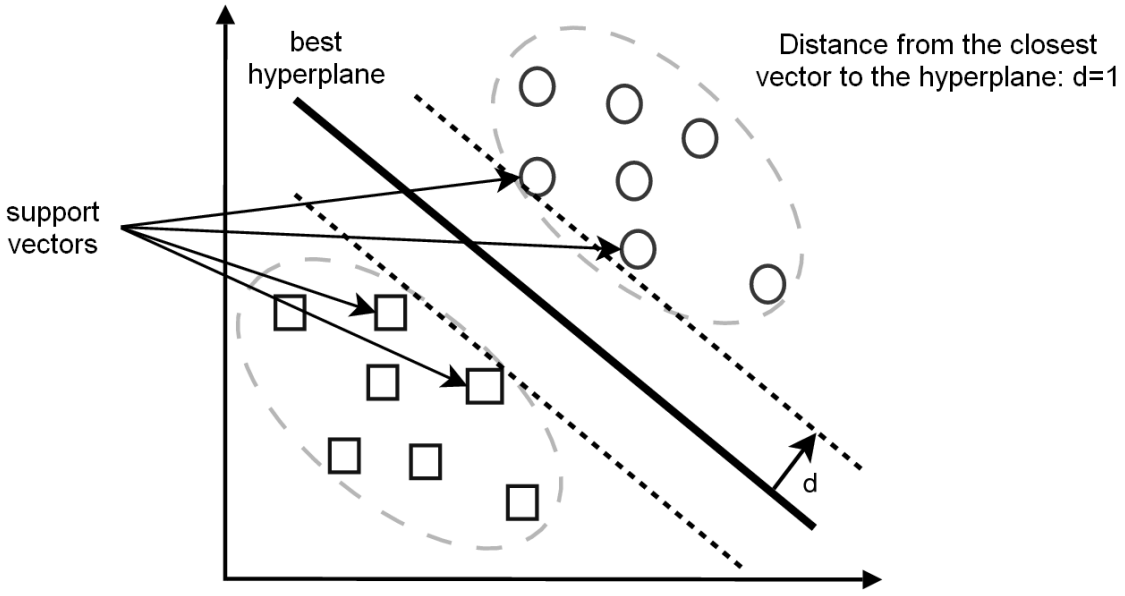

Figure 18 – SVM: best hyperplane

Chapter 5: Results

5.1. "Combination drug therapy" data

The 7 missing values that remain in the real screening dataset after the data treatment presented in Chapter 3: Section 3.3 are imputed. The function "rfImpute" from the package "randomForest" replaces the missing values by a weighted average of the non-missing values.

The models are built using the real screening dataset, with the objective of predicting the target variables "P-gp 15 micromolar efficacy" and "Multifactor usability". Only the features that are common to the real screening dataset and the virtual screening dataset are used to build the models. The data analysis methods are applied in three stages:

- As a regression problem, with the objective of predicting precisely the compounds' efficiency against P-gp. The models' goal is thus to predict the exact continuous value taken by the target variable "P-gp 15 micromolar efficacy";
- As a classification problem, with the objective of predicting if the compounds are effective or not against P-gp. The models' goal is thus to predict a binary value corresponding to "Yes: promising compound" (if the exact value taken by "P-gp 15 micromolar efficacy" is above an arbitrarily chosen threshold) or "No: unpromising compound" (if the exact value is below the threshold);
- As a classification problem, with the objective of predicting the binary target variable "Multifactor usability": a value corresponding to "Yes: promising compound" or "No: unpromising compound".

5.1.1. Regression problem

The advantage of this approach is the precision of the numerical values: having a model that can predict the exact value taken by the target variable "P-gp 15 micromolar efficacy" would mean obtaining results of a good quality.

The regression tree is tuned by varying the complexity parameter from 0 to 0.5. The final tree's complexity parameter is 0.15. The random forest is tuned by varying the number of variables randomly sampled as candidates at each split from 2 to 9. The final random forest tests 2 variables at each split and contains 500 trees. The neural network has one hidden layer. It is tuned by varying the number of nodes in this layer from 1 to 10 and the weight decay from 0 to 0.1. The final neural network's hidden layer contains 7 nodes and uses a weight decay of 0.08. The support vector machine (SVM) is tuned by looking for the best sigma (using the function "sigest" from the package "kernlab") and by varying C (influence of the misclassification). The final SVM uses a Radial Basis Function Kernel and its sigma and C (influence of the misclassifications) are 0.01418318359 and 32 respectively.

The quality of the models is measured by applying them to the dataset they were trained with (the real screening dataset) and computing the RMSE (Root-Mean-Square Error) and the MAPE (Mean Absolute Percentage Error) using the LOOCV method. Table 5 summarizes the results obtained using each of the data analytics methods described in Chapter 4: Section 4.4.

For this regression problem, the model which obtains the best results according to the RMSE and the MAPE is the random forest. This lowest RMSE, 0.14, corresponds to an error of 14%, the target variable "P-gp 15 micromolar efficacy" roughly falling between 0 and 1. The most important predictor variables are available for the each

models: as noted in Table 5, the values taken by the two predictor variables "transition_dbd" and "3b5x_dbd" seem to be the most important to predict the target variable.

Model	RMSE	MAPE	Most important predictors
Linear regression	0.27	6.34 %	transition_nbd, 3b5z_nbd_2, 2hyd_nbd_1, transition_dbd
Stepwise regression	0.27	6.78 %	p450_3a4_eq, 3b5x_dbd, 4ksb_dbd, p450_3a4_uneq, transition_dbd
Regression tree	0.28	3.39 %	3b5x_dbd, 4ksb_dbd, p450_3a4_eq, transition_dbd, p450_3a4_uneq
Random forest	0.14	3.42 %	4ksb_dbd, transition_nbd, transition_dbd, 2hyd_nbd_1, p450_3a4_eq, 3b5x_dbd
Neural network	0.27	6.72 %	transition_nbd, p450_3a4_uneq, 3b5x_dbd, transition_dbd
Support vector machine	0.26	5.34 %	p450_3a4_eq, 3b5x_dbd, 4ksb_dbd, p450_3a4_uneq, transition_dbd

Table 5 – Regression problem: quality of the results obtained by each model

5.1.2. Classification problem

Instead of looking at this problem as a regression problem, this approach consists in predicting the binary target variable "Multifactor usability" and in transforming the continuous target variable "P-gp 15 micromolar efficacy" into a binary variable.

The first step to transform a continuous variable into a binary variable is to choose a threshold value: the value becomes "True" (corresponding to "Yes: this is a promising compound") if the continuous value taken by "P-gp 15 micromolar efficacy" is above this threshold, or "False" (corresponding to "No: this is not a promising compound") if the continuous value is below this threshold. The objective of these new

models is not to determine how efficient a compound is in blocking P-gp, but to determine if it is efficient or inefficient.

The threshold value is either set to 0.7, 0.6 or 0.5. Plotting the continuous target variable "P-gp 15 micromolar efficacy" against the binary target variable "Multifactor usability" (see boxplots in Figure 19) suggests that 0.5 and 0.6 may be better threshold values than 0.7. The No Information Rate for each model is available in Table 6.

Once again, the quality of the models is measured by applying them to the dataset they were trained with (the real screening dataset) and computing the kappa, the accuracy and the area under the curve using the LOOCV method. Table 7 summarizes the results obtained using the data analytics methods described in Chapter 4: Section 4.4. The random forest method was not used in the classification problem because it is overfitting.

As can be observed in Table 7, when the threshold value is set to 0.7, the support vector machine model outperforms all the other models. However, when the threshold value is set to 0.6 or 0.5, the classification tree model outperforms the other models. This is also the case for the models predicting the target variable "Multifactor usability".

Classification model	*Threshold: 0.7*	*Threshold: 0.6*	*Threshold: 0.5*	*Multifactor usability*
Number of compounds classified as efficient in the training dataset (out of 71)	12	21	23	21
No Information Rate	83.10 %	70.42 %	67.61 %	70.42 %

Table 6 – No information rate for each classification problem

	Metric	Logistic regression	Stepwise regression	Classification tree	Neural network	Support vector machine
Threshold: 0.7	**Kappa**	0.45	0.25	0.39	0.13	0.89
	Accuracy	88.73	85.92	78.87	84.51	97.18
	Area under the curve	0.67	0.58	0.74	0.54	0.92
Threshold: 0.6	**Kappa**	0.30	0.19	0.63	0.25	0.30
	Accuracy	76.06	74.65	85.92	85.92	77.46
	Area under the curve	0.62	0.57	0.79	0.58	0.62
Threshold: 0.5	**Kappa**	0.18	0.22	0.61	0.19	0.51
	Accuracy	69.01	73.24	81.69	71.83	81.69
	Area under the curve	0.58	0.59	0.83	0.58	0.72
"Multifactor usability"	**Kappa**	0.24	0.29	0.59	0.30	0.38
	Accuracy	85.92	72.73	83.12	75.32	79.22
	Area under the curve	0.58	0.63	0.79	0.62	0.65

Table 7 – Classification problem: quality of the results obtained by each model

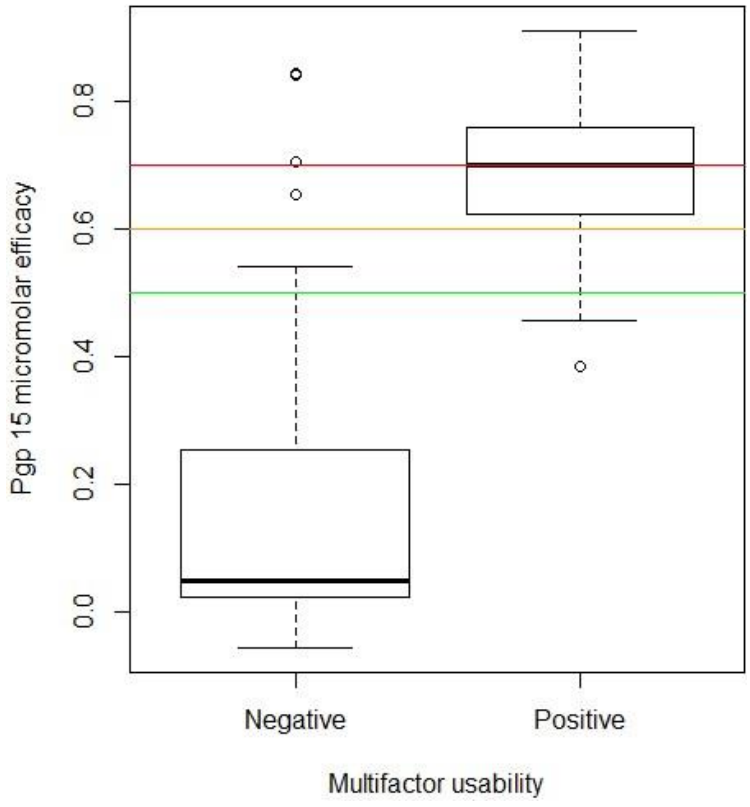

Figure 19 – Relationship between the two target variables

5.2. Selection of potentially promising compounds

The predictive models are applied to the whole real screening dataset (no LOOCV) in order to obtain the final model. The regression models and the classification models are then applied to the virtual screening dataset separately in order to find new potentially promising compounds to purchase. The first step is to impute the missing values in the virtual screening dataset. This is done by mean imputation: each missing value is replaced by the mean of the non-missing values taken by the variable.

5.2.1. Application of the regression models

The complexity parameter of the final tree (depicted in Figure 20) is 0.15. The final random forest tests 2 variables at each split and contains 500 trees. The final neural network (depicted in Figure 21) has 1 hidden layer composed of 5 nodes and uses a weight decay of 0.09. The final SVM uses a Radial Basis Function Kernel and its epsilon, sigma and C (influence of the misclassifications) are 0.1, 0.01642210808 and 16 respectively. It is composed of 66 support vectors. An excerpt of the results obtained by each regression model is presented in Table 8.

When an incorrect value is predicted either by the linear regression model or by the stepwise regression model, the compound is automatically rejected. Incorrect values are defined as values which are lower than -0.5 or higher than 1.5, since the continuous target variable “P-gp 15 micromolar efficacy” is between 0 and 1.

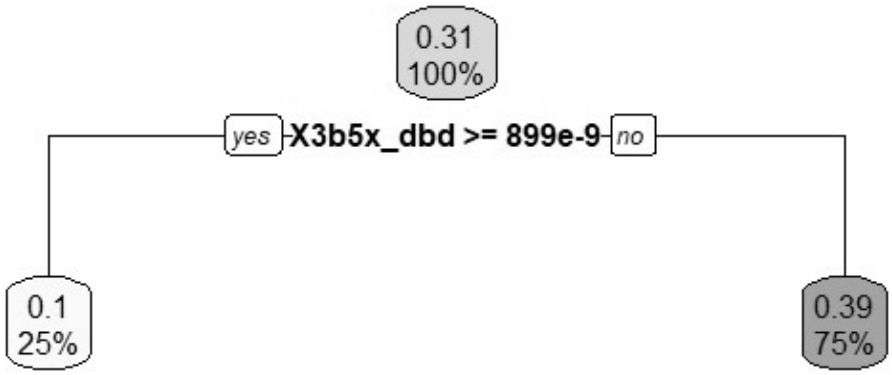

Figure 20 – Final model: regression tree

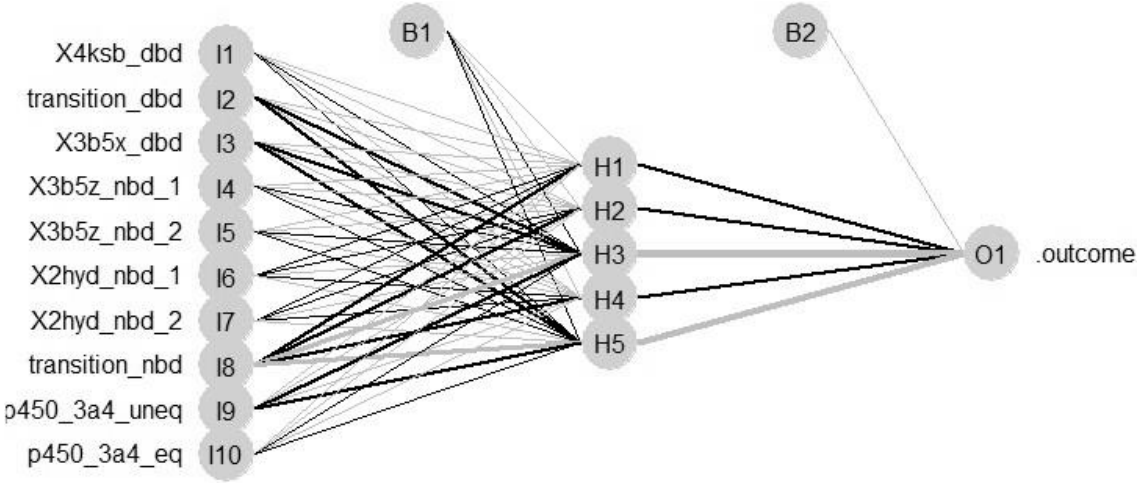

Figure 21 – Final model: regression neural network

Compound	*Linear reg.*	*Stepwise reg.*	*Regres. tree*	*Random forest*	*Neural network*	*SVM*
ZINC00000017	0.211	-0.686	0.103	0.187	0.115	0.367
ZINC00000190	1.797	0.814	0.103	0.177	0.129	0.370
ZINC00000357	1.742	-1.013	0.103	0.124	0.115	0.370
ZINC00000534	-2.506	-0.979	0.103	0.114	0.115	0.370
ZINC00000757	0.681	-7.811	0.103	0.179	0.115	0.370
ZINC00000842	-1.209	-0.577	0.103	0.181	0.116	0.370
ZINC00000868	6.936	2.816	0.103	0.179	0.115	0.370
ZINC00000982	8.098	-1.006	0.103	0.179	0.115	0.370
ZINC00001098	1.137	-1.516	0.103	0.134	0.115	0.370
ZINC00001239	4.455	1.339	0.103	0.179	0.115	0.370
ZINC00001282	1.741	-0.486	0.103	0.177	0.117	0.365
ZINC00001380	1.329	-1.786	0.103	0.179	0.115	0.370
ZINC00001402	-0.771	-1.182	0.386	0.285	0.118	0.329
ZINC00001524	4.275	-1.421	0.103	0.179	0.115	0.370
ZINC00001734	-2.212	-0.775	0.103	0.180	0.115	0.370

Table 8 – Excerpt: efficiencies predicted by the regression models (rounded values)

Compound	*Logistic reg.*	*Stepwise reg.*	*Classif. tree*	*Neural network*	*SVM*
ZINC00000017	FALSE	FALSE	FALSE	FALSE	FALSE
ZINC00000190	TRUE	FALSE	FALSE	FALSE	FALSE
ZINC00000357	FALSE	FALSE	FALSE	FALSE	FALSE
ZINC00000534	FALSE	FALSE	FALSE	FALSE	FALSE
ZINC00000757	FALSE	FALSE	FALSE	FALSE	FALSE
ZINC00000842	TRUE	FALSE	FALSE	FALSE	FALSE
ZINC00000868	FALSE	FALSE	FALSE	FALSE	FALSE
ZINC00000982	FALSE	FALSE	FALSE	FALSE	FALSE
ZINC00001098	FALSE	FALSE	FALSE	FALSE	FALSE
ZINC00001239	FALSE	FALSE	FALSE	FALSE	FALSE
ZINC00001282	FALSE	FALSE	FALSE	FALSE	FALSE
ZINC00001380	FALSE	FALSE	FALSE	FALSE	FALSE
ZINC00001402	FALSE	FALSE	TRUE	FALSE	FALSE
ZINC00001524	FALSE	FALSE	FALSE	FALSE	FALSE
ZINC00001734	FALSE	FALSE	FALSE	FALSE	FALSE

Table 9 – Excerpt: efficiencies predicted by the Multifactor usability models

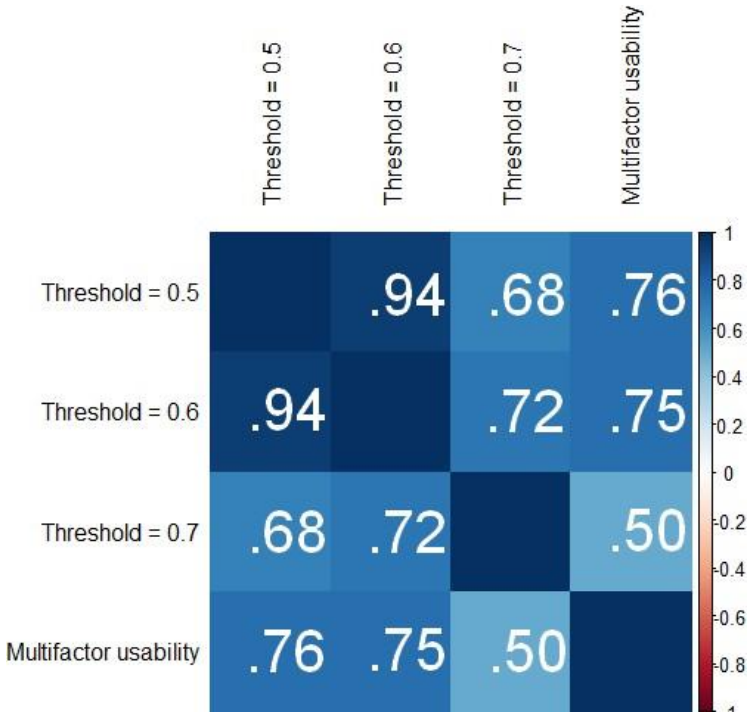

Figure 22 – Correlations between the values taken by the binary targets

5.2.2. Application of the classification models

An excerpt of the results obtained by each classification model whose target is the variable “Multifactor usability” is presented in Table 9.

As can be observed in Figure 22, the real values taken by the target variables when the threshold value is 0.5 and when the threshold value is 0.6 are highly correlated (94%). In 75% of cases, the variable “Multifactor usability” agrees with them. However, when the threshold value is 0.7, the variable “Multifactor usability” only agrees in 50% of cases. For this reason, the models whose objective is to predict the binary variable with a threshold of 0.7 are not used in the final model.

The three classification tree models obtained by using a threshold value of 0.5 on the continuous target variable “P-gp 15 micromolar efficacy”, by using a threshold value of 0.6 and by predicting the binary target variable “Multifactor usability” perform the best on the training dataset (see Table 7). These trees are depicted in Figure 23, Figure 24 and Figure 25 an excerpt of their predictions is presented in Table 10. When less than two of these three models classify a compound as “Promising”, the compound is automatically rejected.

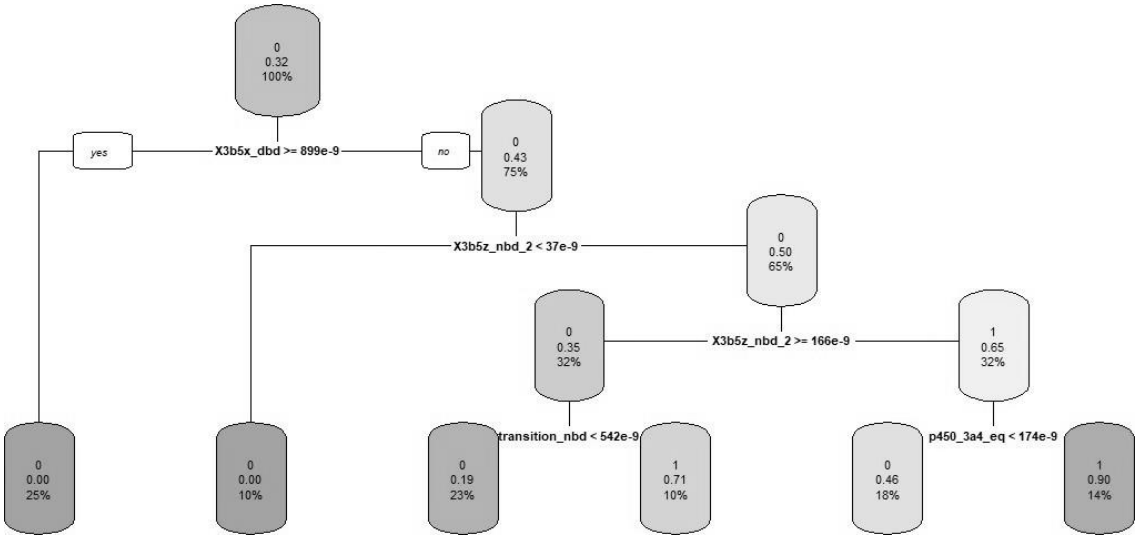

Figure 23 – Final model: classification tree (threshold = 0.5)

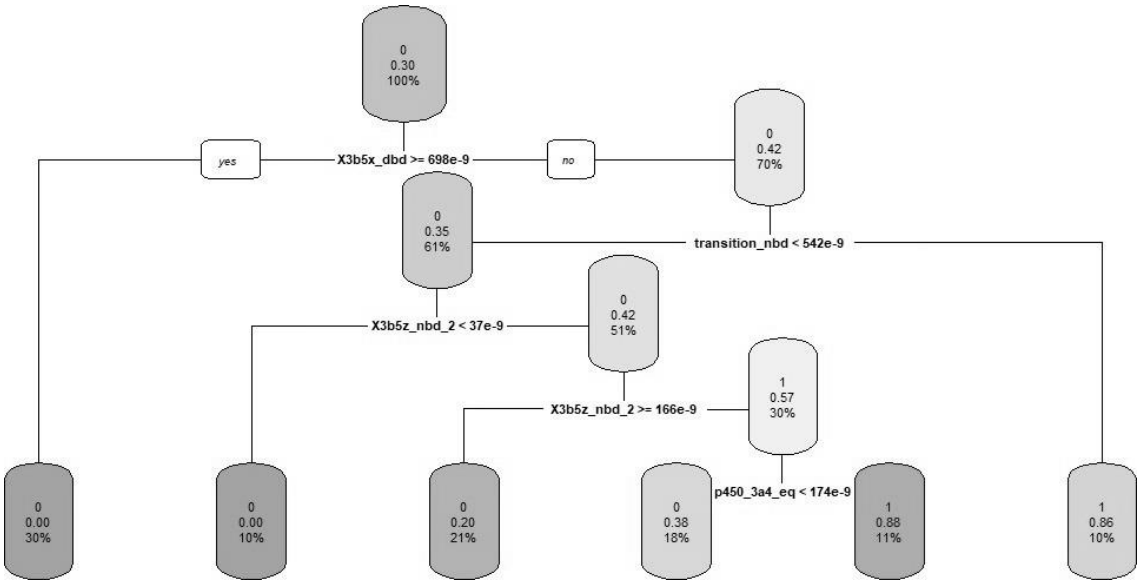

Figure 24 – Final model: classification tree (threshold = 0.6)

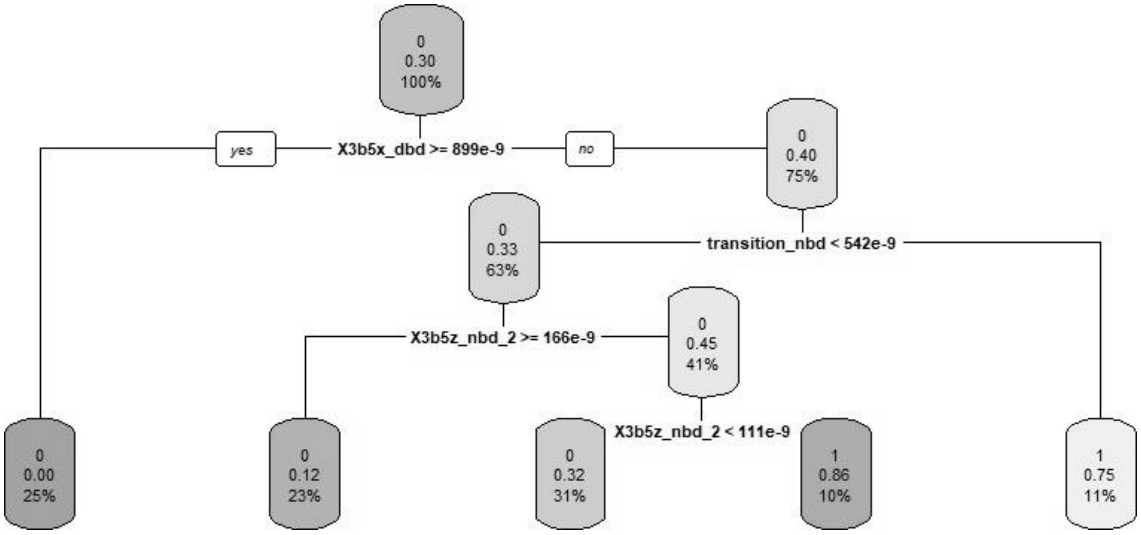

Figure 25 – Final model: classification tree (Multifactor usability)

Compound	*Threshold 0.5*	*Threshold 0.6*	*"Multifactor usability"*
ZINC00001282	FALSE	FALSE	FALSE
ZINC00001380	FALSE	FALSE	FALSE
ZINC00001402	TRUE	FALSE	TRUE
ZINC00001524	FALSE	FALSE	FALSE
ZINC00002646	FALSE	FALSE	FALSE
ZINC00003704	FALSE	FALSE	FALSE
ZINC00003837	FALSE	FALSE	FALSE
ZINC00006296	FALSE	FALSE	FALSE
ZINC00006662	TRUE	TRUE	TRUE
ZINC00027763	FALSE	FALSE	FALSE
ZINC00027913	TRUE	TRUE	TRUE
ZINC00034310	FALSE	FALSE	FALSE
ZINC00034317	FALSE	FALSE	FALSE
ZINC00034348	FALSE	FALSE	FALSE
ZINC00034379	TRUE	FALSE	TRUE

Table 10 – Excerpt: efficiencies predicted by the classification trees

5.2.3. Final model and selection of promising compounds

Six values are predicted by the six regression models (see Chapter 5: Section 5.1.1): the linear regression, the stepwise regression, the regression tree, the random forest, the neural network and the support vector machine. The predicted efficiency of the compound against P-gp is defined as the mean of these six values. Figure 26 illustrates this process:

- The predicted efficiency is the mean of the continuous predictions made by the regression models;
- The selected classification models validate or invalidate this predicted efficiency.

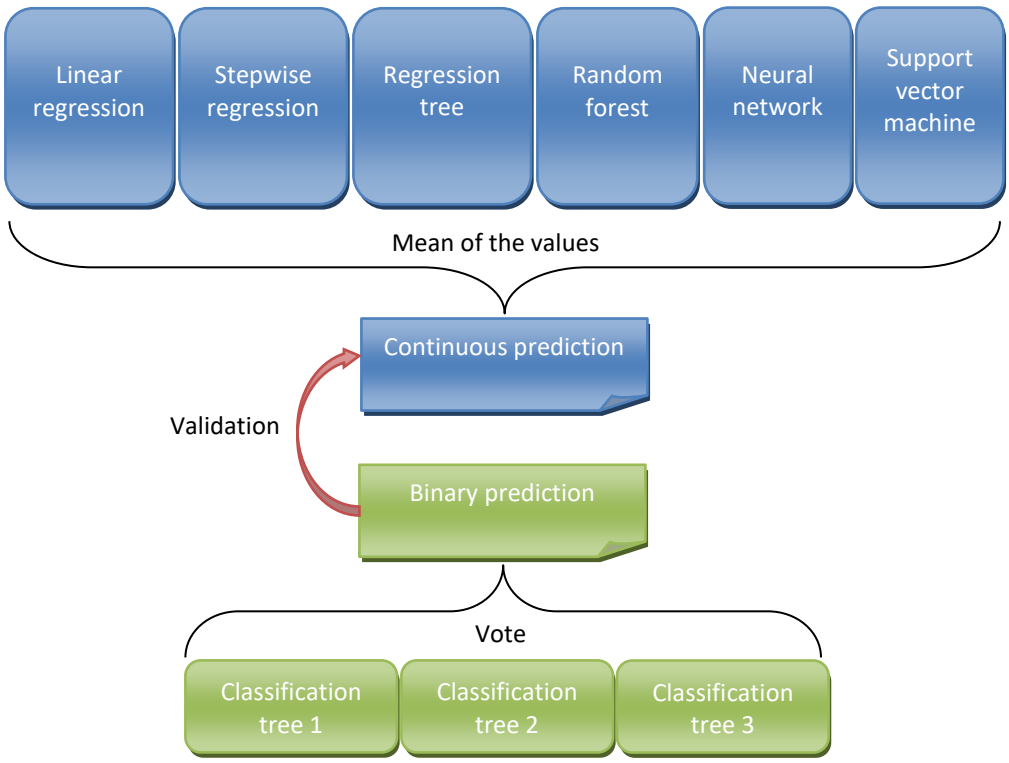

Figure 26 – Use of the models to identify promising compounds

This process is applied to the real screening dataset to measure the quality of the predictions. An excerpt of the predicted values can be seen in Table 11. This final model obtains an RMSE of 0.24 and a MAPE of 5.68% for the regression problem and a kappa of 0.58, an accuracy of 81.69 and an AUC of 0.80 for the classification problem. The predicted values can be compared to the real values for the regression problem and the classification problem in Figure 27 and Table 12 respectively.

Compound	*P-gp 15 micro. efficacy*	*P-gp 15 micro. efficacy prediction*	*Multifactor usability*	*Multifactor usability prediction*
ZINC23175200	0.189	0.340	FALSE	FALSE
ZINC9446321	0.009	0.262	FALSE	FALSE
ZINC84559953	0.839	0.687	TRUE	TRUE
ZINC23337273	0.506	0.355	TRUE	TRUE
ZINC4741719	0.041	0.264	FALSE	FALSE

Table 11 – Excerpt: values predicted by the final model (rounded values)

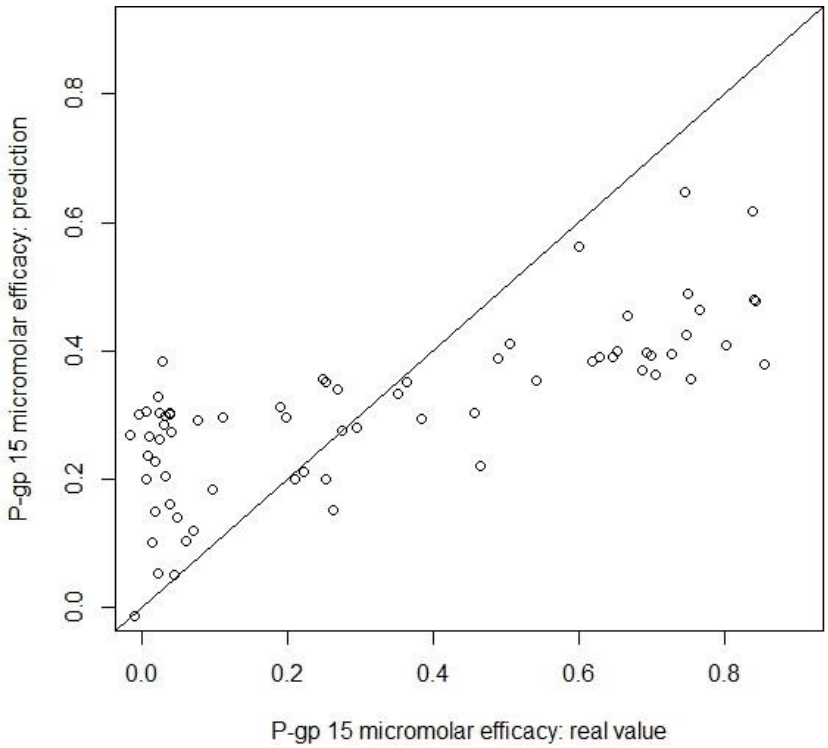

Figure 27 – Regression problem: predicted values versus real values

		Multifactor usability: real value	
		FALSE	**TRUE**
Multifactor usability: predicted value	**FALSE**	42	8
	TRUE	5	16

Table 12 – Classification problem: confusion matrix

For each model, the most important predictors are identified using the function "varImp" from the package "caret". This data is available in Table 13 for the regression models and in Table 14 for the classification models. The predictors "4ksb_dbd", "3b5x_dbd", "transition_dbd" and "transition_nbd" are often important in the models' design.

Predictor	*Linear regression*	*AIC stepwise regression*	*Regression tree*	*Random forest*	*Neural network*	*Support vector machine*
4ksb_dbd	12.32	**83.31**	**72.29**	**80.79**	17.14	**83.31**
transition_dbd	25.77	**64.31**	**71.87**	**100.00**	51.85	**64.31**
3b5x_dbd	0.00	**89.83**	**71.94**	**63.78**	**60.24**	**89.83**
3b5z_nbd_1	3.07	0.00	100.00	0.00	14.02	0.00
3b5z_nbd_2	73.01	38.38	78.30	54.10	33.27	38.38
2hyd_nbd_1	33.19	34.96	0.00	68.41	3.54	34.96
2hyd_nbd_2	24.77	40.50	0.00	27.77	0.00	40.50
transition_nbd	**100.00**	24.06	0.00	**90.49**	**100.00**	24.06
p450_3a4_uneq	18.33	75.77	0.00	58.79	79.96	75.77
p450_3a4_eq	1.84	100.00	0.00	61.53	29.61	100.00

Table 13 – Importance of each predictor for each regression model

Predictor	*Threshold 0.5*	*Threshold 0.6*	*Multifactor usability*
p450_3a4_eq	10.26	8.18	4.70

p450_3a4_uneq	4.33	5.36	4.93
transition_dbd	4.66	7.52	7.37
transition_nbd	5.93	6.33	5.85
2hyd_nbd_1	4.77	6.05	1.13
2hyd_nbd_2	1.80	3.38	4.35
3b5x_dbd	7.28	5.22	5.60
3b5z_nbd_1	0.77	0.00	1.28
3b5z_nbd_2	6.01	7.95	5.24
4ksb_dbd	6.98	6.04	3.99

Table 14 – Importance of each predictor for each classification model

The final model is applied to the virtual screening dataset. The potentially promising compounds are selected by sorting the list of compounds by predicted efficiency. The output of this process is a list of the 50 most promising compounds: those whose predicted efficiency is the highest. The 5 first compounds and their predicted efficiency in blocking P-gp are listed in Table 15.

Compound	*Predicted efficiency*
ZINC58162043	0.845
ZINC65099984	0.704
ZINC12190220	0.692
ZINC12321091	0.685
ZINC13724176	0.680

Table 15 – Top 5 potentially promising compounds (rounded values)

5.3. “AutoDock Vina” data

As developed in Chapter 2: Section 2.2, the objective is to select *for* molecules that interact preferentially with the Nucleotide Binding Domains and to select *against* molecules that interact preferentially with the Drug Binding Domain. As developed in Chapter 3: Section 3.2.2, Both the Kd and the dG are indicators of the strength of the interaction between a compound and a receptor on P-gp.

By comparing the Kd and the dG, the fact that these two variables are closely related is brought to light. In Figure 28, each Kd and dG of the AutoDock Vina results dataset is plotted by dots. The exponential trend line proves that the variables are dependent. The Kd is discarded to avoid keeping duplicates of the same information.

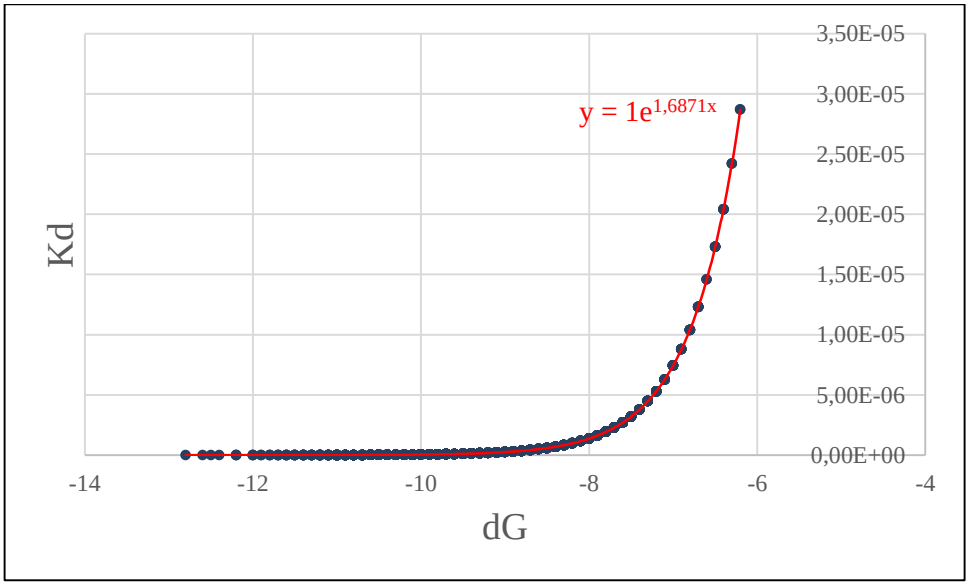

Figure 28 – Relationship between Kd and dG

5.3.1. Variance of the center of mass

I study the variance in the center of mass of the compounds when they dock to P-gp. The idea is that this information can help determine if the molecule docks to the same spot of P-gp repeatedly or if the docking spot is highly variable.

As developed in Chapter 3: Section 3.2.2, the coordinates (x, y and z) of each atom that make up the compound when it is docked to a receptor on P-gp are available. This information is used to calculate the coordinates of the center of mass for each compound-receptor couple, for each of the 20 docking positions. I proceed as follows:

- **Step 1:** each atom is associated to its atomic mass:

Element	Boron	Carbon	Chlorine	Fluorine	Hydrogen	Nitrogen	Oxygen	Sulfur
Symbol	B	C	Cl	F	H	N	O	S
Atomic mass	10.811	12.0107	35.453	18.998403	1.007940	14.0067	15.999	32.065

Table 16 – Molar masses of some chemical elements

- **Step 2:** the three coordinates are calculated using the following formulas:

$$COMx = \frac{SUMPRODUCT(x\ coordinates, molar\ masses)}{SUM(molar\ masses)}$$

$$COMy = \frac{SUMPRODUCT(y\ coordinates, molar\ masses)}{SUM(molar\ masses)}$$

$$COMz = \frac{SUMPRODUCT(z\ coordinates, molar\ masses)}{SUM(molar\ masses)}$$

An example of the result of such a calculation is presented in Table 17. An excerpt of the dataset containing the coordinates of the center of mass for each position, for each compound-receptor couple is presented in Table 18.

Atom	*Molar mass*	*x*	*y*	*z*
N	14.0067	-7.821	3.596	26.281
C	12.0107	-7.114	4.716	26.913
N	14.0067	-6.033	4.201	27.768
C	12.0107	-5.471	3.089	27.433
N	14.0067	-5.81	2.312	26.328
C	12.0107	-5.122	1.194	26.108
O	15.999	-5.367	0.481	25.155
C	12.0107	-4.043	0.847	27.068
C	12.0107	-3.193	-0.243	27.076
C	12.0107	-3.256	-1.439	26.209
C	12.0107	-4.118	-1.832	25.188
C	12.0107	-3.915	-3.039	24.549
C	12.0107	-2.858	-3.854	24.923
C	12.0107	-2	-3.473	25.935
C	12.0107	-2.187	-2.262	26.59
N	14.0067	-1.484	-1.651	27.623
C	12.0107	-2.021	-0.463	27.946
O	15.999	-1.616	0.297	28.806
Cl	35.453	-4.984	-3.536	23.273
S	32.065	-4.141	2.217	28.212
C	12.0107	-6.925	2.773	25.456
H	12.0107	-0.701	-2.031	28.051
C	12.0107	-8.914	4.05	25.535
C	12.0107	-10.073	3.288	25.465
C	12.0107	-11.155	3.737	24.732
C	12.0107	-11.082	4.944	24.056
O	15.999	-11.992	5.597	23.271
C	12.0107	-9.92	5.709	24.12
C	12.0107	-11.525	6.956	23.192
O	15.999	-10.101	6.84	23.376
C	12.0107	-8.839	5.261	24.858
Center of mass:		-5,8335521	1,3895055	25,8198869

Table 17 – COM coordinates for “ZINC783138” docking to “2hyd_1_cyt”, position 1

Compound	*Receptor*	*Docking position*	*x*	*y*	*z*
ZINC00601275_SMU113	2hyd_dbd_1	1	-10.5359351	-13.61404154	6.2096014
ZINC00601275_SMU113	2hyd_dbd_1	2	1.2523774	-7.80591368	2.1742224
ZINC00601275_SMU113	2hyd_dbd_1	3	-10.9861816	-13.70722547	5.7324860
ZINC00601275_SMU113	2hyd_dbd_1	4	-10.4461159	-14.29629766	6.5037984
ZINC00601275_SMU113	2hyd_dbd_1	5	-10.2746077	-14.03233708	6.3522479
ZINC00601275_SMU113	2hyd_dbd_1	…	…	…	…
ZINC00601275_SMU113	2hyd_dbd_1	20	-10.5774712	-13.84897089	6.8505924
ZINC00601275_SMU113	2hyd_dbd_2	1	-8.8125101	-14.10053003	6.2111602
ZINC00601275_SMU113	2hyd_dbd_2	…	…	…	…

Table 18 – Excerpt: "Coordinates of the center of mass" dataset

- **Step 3:** the variance of each coordinate over the 20 positions is calculated.

An excerpt of the dataset containing the results of this calculation is presented in Table 19.

Compound	*Receptor*	*Variance of x*	*Variance of y*	*Variance of z*
ZINC00601275_SMU113	2hyd_dbd_1	25.2957012	6.2500934	3.3964526
ZINC00601275_SMU113	2hyd_dbd_2	51.9882874	137.2674703	30.4689793
…	…	…	…	…

Table 19 – Excerpt: "Variance of the coordinates of the center of mass" dataset

- **Step 4:** finally, for each compound-receptor couple, the mean of the variances of the three coordinates is calculated so that the dataset can be ordered by increasing variance of the center of mass.

The threshold under which a variance is considered low is arbitrarily set to 5. The variance of the center of mass is considered low if all three variances of the coordinates are low. Out of the 1921 compound-receptor couples, 66 have a low center of mass variance. In these 66 couples, 47 out of the 113 different compounds and 5 out of the 17

different receptors are represented. The receptors “3b5x_dbd” and “4ksb_dbd” are especially well represented: these two spots are creating favorable biochemical interactions. The other three receptors which are represented are “3b5z_dbd_1”, “transition_2_dbd” and “mdr1_6_nbd”. The only represented “NBD” receptor is therefore located on MDR1_6, the Multi-Drug Resistance Protein 1.

The 5 compound-receptor couples with the lowest average variance are presented in Table 20.

Compound	*Receptor*	*Variance x*	*Variance y*	*Variance z*	*Variance average*	*Kd*
ZINC07006681 SMU95	3b5x_dbd	0.286	0.215	0.219	0.240	-7.9
ZINC84559953 SMU96	3b5x_dbd	0.388	0.180	0.328	0.299	-8.7
ZINC41469020 SMU110	3b5z_dbd_1	0.167	0.573	0.166	0.302	-8.2
ZINC33326619 SMU94	3b5x_dbd	0.296	0.428	0.267	0.330	-8.4
ZINC12577459 SMU97	3b5x_dbd	0.415	0.404	0.257	0.359	-8.4

Table 20 – Compound-Receptor couples with the lowest center of mass variance

In the case of P-gp, the variance of the center of mass is only low for compound-receptors couples with a receptor located on the Drug Binding Domain. If the hypothesis stated in Chapter 3: Section 3.2.2 is correct (a compound that docks repeatedly to the same spot on the DBD is more easily transported out of the cell, whereas a compound that docks repeatedly to the same spot on the NBD is more efficient against P-gp), these compounds may therefore be inadequate for P-gp inhibition.

5.3.2. Relationships between the AutoDock Vina data and the Combination drug therapy data

The compounds have names of the form “ZINC123” in the real screening dataset and names of the form “ZINC123 SMU456” in the AutoDock Vina dataset (see Table 2 and Table 4). I merge these two datasets by ignoring the “SMU” identification number.

Out of the 77 compounds from the AutoDock Vina dataset, 63 are present as well in the real screening dataset. However, one of these compounds, “ZINC9224466” in the real screening dataset, has to be disregarded because it matches with both “ZINC09224466_SMU32” and “ZINC09224466_SMU102” in the AutoDock Vina dataset and there is no way of knowing which of the two refers to this compound.

The final merged dataset, which contains the remaining 62 common compounds, is used to study the relationships between variables. For the receptors that are common to the two datasets, the binding affinity (from the real screening dataset) and the Kd (from the AutoDock Vina dataset) are correlated (see the example of the receptor 3b5x_dbd in Figure 29). However, the binding affinity and the variance of the center of mass are not.

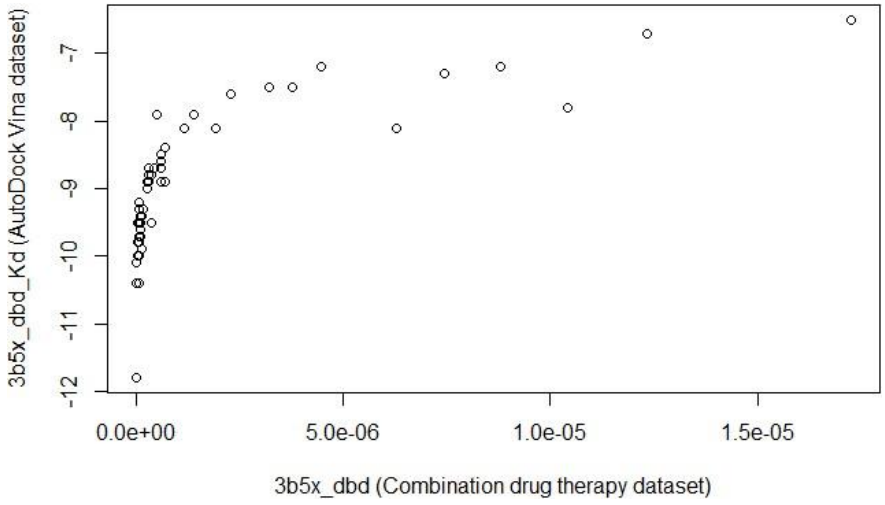

Figure 29 – Relationship between the binding affinity and the Kd for the receptor 3b5x_dbd

The merged dataset allows the study the relationships between target variables. The target variables are “P-gp 15 micromolar efficacy” and “Multifactor usability” from the real screening dataset, and the variance of the center of mass and the Kd from the AutoDock Vina dataset. Figure 30 illustrates the relationships between these variables with a heat map:

The continuous variable “P-gp 15 micromolar efficacy” and the binary variable “Multifactor usability” are highly correlated, which is expected. However, there is no clear correlation between the targets “P-gp 15 micromolar efficacy” or “Multifactor usability” and the various targets from the AutoDock Vina dataset.

All the Kd variables are positively correlated to each other because the original Kd variable only takes negative values. There are no clear correlations between the Kd of a receptor and the variance of its center of mass.

The variance of the center of masse of the receptor “3b5z_nbd_2” is the only one that is correlated to the variance of the center of mass of other receptors. It is, for example, negatively correlated to the variance of the center of mass of “3b5x_dbd” and positively correlated to the variance of the center of mass of “transition_1_dbd”, which means that if a compound docks repeatedly to “3b5z_nbd_2”, it is likely that it will dock to “transition_1_dbd” too and unlikely that it will dock to “3b5x_dbd”.

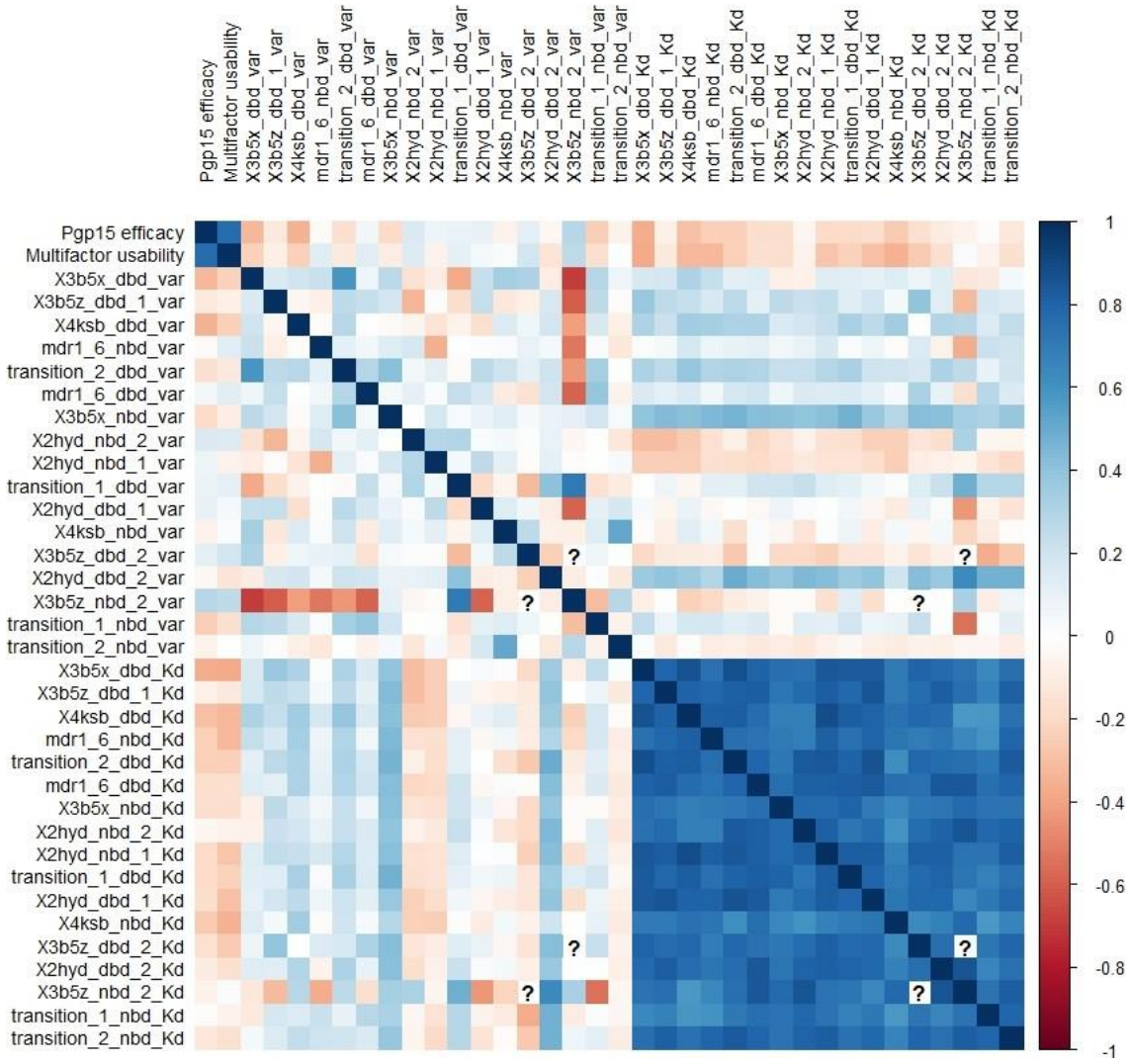

Figure 30 – Correlations between the target variables

5.4. Validation

Few of the compounds selected as promising are present in the other datasets. This comes from the fact that these compounds are selected out of the 159,000 compounds from the virtual screening dataset, which is the biggest dataset: the real screening dataset only contains 77 compounds and the AutoDock Vina dataset only contains 123 compounds.

One of the 50 compounds that are identified by the models as potentially promising compounds (see Chapter 5: Section 5.2.3 and Table 15) is present in the real screening dataset: this means that this compound was selected by the experts, when they chose the original group of 31 compounds to purchase from the original virtual screening database containing 159,000 compounds. This compound is referred to as "ZINC84559953". Its predicted efficiency is 0.634 and its real efficiency, which is known thanks to the real screening dataset, is 0.839. According to the target variable "Multifactor usability", it is indeed effective against P-gp.

Out of the 62 compounds common to both the real screening dataset and the AutoDock Vina dataset, 31 have a low variance of the center of mass when they dock to one or more DBD receptors (44 compound-receptor couples in total). An excerpt of this data is available in Table 21. If the hypothesis stated in Chapter 3: Section 3.2.2 is correct (a compound that docks repeatedly to the same spot on the DBD is more easily transported out of the cell, whereas a compound that docks repeatedly to the same spot on the NBD is more efficient against P-gp), these compounds are inadequate for P-gp inhibition.

Out of these 31 compounds, 10 compounds are classified as promising by the target variable “Multifactor usability”, which contradicts the hypothesis. However, the same percentage (30%) of all the compounds in the real screening dataset are classified as promising by the target variable “Multifactor usability” (see Table 6): this suggests that the variance of the center of mass does not have any impact on the efficiency against P-gp. Conducting further analysis on more compounds in the future may provide more insights and allow to confidently confirm or refute this hypothesis.

Compound	*Receptor*	*Variance of the center of mass*	*Multifactor usability*
ZINC23175200	3b5x_dbd_ori	0.401	FALSE
ZINC23175200	4ksb_dbd_ori	0.624	FALSE
ZINC25220272	4ksb_dbd_ori	1.183	FALSE
ZINC25376436	4ksb_dbd_ori	0.742	TRUE
ZINC12783661	3b5x_dbd_ori	0.571	FALSE

Table 21 – Excerpt: Compounds binding to the DBD with a low variance of COM

Chapter 6: Conclusion and future work

By temporarily turning off P-gp transporters, the multidrug resistance of cancerous cells can be reversed and they can become sensitive again to chemotherapeutic drugs. The primary objective of this work is to look for molecules which have the ability of turning off P-gp.

Analyzing molecules to determine their efficiency against P-gp can be done using expensive processes, such as the High-Throughput Screening of an entire library of chemical compounds. By using virtual screening software such as AutoDock Vina and machine learning methods to predict the efficiency of the molecules, a lot of money can be saved by the researchers.

The results obtained in this work show that predicting the efficiency of a chemical compound is complex. The random forest model proves to be the most efficient in the case of the regression problem and the classification tree models prove to be the most efficient in the case of the classification problem. The potentially promising compounds, which are selected after applying data analytics models, will have to be purchased and analyzed by CD4 before I know if they are actually efficient. Analyzing the compounds will also allow the accuracy of the final model to be measured.

Some key characteristics of promising compounds for P-gp inhibition have been identified: high values for the predictor variables “3b5x_dbd”, “4ksb_dbd” and “transition_dbd”, which indicate a weak binding to these DBD receptors, and low values for the predictor variable “transition_nbd”, which indicates a strong binding to this NBD receptor. I made the hypothesis that a low variance of the center of mass of a

compound when it docks to specific receptors on P-gp could be a good indicator of the efficiency of the compound against P-gp. In this work, this hypothesis does not seem true. However, it needs further analysis to be confidently confirmed or refuted.

The main issue encountered in this work is the fact that many values are missing from the real screening dataset, which is the training dataset, and that this dataset does not contain enough samples to build robust prediction models.

The objective is to find compounds that sensitize the cell so that the chemotherapeutic can kill it, not compounds that kill the cell themselves. Used at the concentration where they reverse multidrug resistance, the compounds should be harmless to non-cancerous cells. As detailed in Chapter 3: Section 3.1.2, the target variable "Actual toxicity" should not be higher than 0.5 for a compound to be selected. This is not taken into account in this work because it requires the building of a different model, capable of predicting the target variable "Actual toxicity".

The target variable "10 micromolar P450 inhibition" should be taken into account in the future as well: a model capable of predicting this variable would allow the selection of compounds that have as few undesirable side effects on the patient as possible.

References

Belle, A., Thiagarajan, R., Reza Soroushmehr, S. M., Navidi, F., Beard, D. A., & Najarian, K. (2015). Big Data Analytics in Healthcare. *BioMed Research International, 2015*, Article ID 370194.

Bikadi, Z., Hazai, I., Malik D, Jemnitz, K., Veres, Z., Hari, P., . . . Mao, Q. (2011). Predicting P-Glycoprotein-Mediated Drug Transport Based On Support Vector Machine and Three-Dimensional Crystal Structure of P-glycoprotein. *PLoS One, 6*(10), e25815.

Brewer, F. K. e. a. (2014). In Vitro Screening for Inhibitors of P-Glycoprotein That Target the Nucleotide Binding Domains. *Molecular Pharmacology, 86.6*, 716–726.

Chapron, A., & Larin, O. (2004-2016). Introduction to Drug Discovery. *Combinatorial Chemistry Review*.

Coates, J., Souhami, L., & El Naqa, I. (2016). Big Data Analytics for Prostate Radiotherapy. *Frontiers in Oncology, 6*, Article number 149.

Dolghih, E., Bryant, C., Renslo, A. R., & Jacobson, M. P. (2011). Predicting Binding to P-Glycoprotein by Flexible Receptor Docking. *PLoS Computational Biology, 7*, e1002083.

El Naqa, I. (2016). Perspectives on making big data analytics work for oncology. *Methods, 111*, 32-44.

Follit, C. A., Brewer, F. K., Wise, J. G., & Vogel, P. (2015). In Vitro Identified Targeted Inhibitors of P-Glycoprotein Overcome Multidrug Resistance in Human Cancer Cells in Culture. *Pharmacology Research & Perspectives, 3.5*.

High Throughput Screening: Methods and Protocols. (2016). (W. P. Janzen Ed. 3 ed.).

Housman, G., Byler, S., Heerboth, S., Lapinska, K., Longacre, M., Snyder, N., & Sarkar, S. (2014). Drug Resistance in Cancer: An Overview. *Cancers, 6*, 1769-1792.

Hughes, J., Rees, S., Kalindjian, S., & Philpott, K. (2011). Principles of early drug discovery. *British Journal of Pharmacology, 162*, 1239-1249.

Kitchen, D., Decornez, H., Furr, J., & Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature Reviews. Drug Discovery, 3*, 935-949.

Li, W., Zhang, H., Assaraf, Y. G., Zhaod, K., Xue, X., Xie, J., . . . Chen, Z.-S. (2016). Overcoming ABC transporter-mediated multidrug resistance: Molecular

mechanisms and novel therapeutic drug strategies. *Drug Resistance Updates*, *27*, 14-29.

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

Luqmani, Y. A. (2005). Mechanisms of Drug Resistance in Cancer Chemotherapy. *Medical Principles and Practice, 14*, 35-48.

McCormick, J. W., Vogel, P. D., & Wise, J. G. (2015). Multiple Drug Transport Pathways through Human P-Glycoprotein. *Biochemistry, 54*, 4374-4390.

McGuire, S. (2014). World Cancer Report 2014: W.H. Organization.

Ow, G. S., & Kuznetsov, V. A. (2016). Big genomics and clinical data analytics strategies for precision cancer prognosis. *Scientific Reports, 6*, Article number 36493

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health Information Science and Systems, 2*.

Robey, R. W., Shukla, S., Finley, E. M., Oldham, R. K., Barnett, D., Ambudkar, S. V., . . . Bates, S. E. (2008). Inhibition of P-glycoprotein (ABCB1)- and multidrug resistance-associated protein 1 (ABCC1)-mediated transport by the orally administered inhibitor, CBT-1((R)). *Biochem Pharmacol, 75*, 1302-1312.

Trott, O., & Olson, A. J. (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *Journal of Computational Chemistry, 31*(2), 455-461.

Zhang, Z. (2016). Variable selection with stepwise and best subset approaches. *Annals of Translational Medicine, 4*(7), 136.

Appendix A: Variables available in the virtual screening dataset

Variable name	Type	Percentage of missing values
COMPOUND	factor	0%
X4ksb_dbd	numeric	1%
transition_dbd	numeric	24%
X3b5x_dbd	numeric	35%
X3b5z_nbd_1	numeric	48%
X3b5z_nbd_2	numeric	36%
X2hyd_nbd_1	numeric	7%
X2hyd_nbd_2	numeric	55%
transition_nbd	numeric	0%
X2hyd_nbd_dbd	numeric	1%
p450_3a4_uneq	numeric	0%
p450_3a4_eq	numeric	21%
bcrp_nbd_1	numeric	34%
msba_dbd	numeric	15%
msba_nbd_1	numeric	23%
msba_nbd_2	numeric	16%

Appendix B: Variables available in the real screening dataset

Variable name	Type	Percentage of missing values
COMPOUND	factor	0%
X4ksb_dbd	numeric	3%
transition_dbd	numeric	3%
X3b5x_dbd	numeric	3%
X3b5z_nbd_1	numeric	4%
X3b5z_nbd_2	numeric	3%
X2hyd_nbd_1	numeric	4%
X2hyd_nbd_2	numeric	3%
transition_nbd	numeric	3%
X2HYD_NBD_DBD		100%
p450_3a4_uneq	numeric	3%
p450_3a4_eq	numeric	9%
bcrp_nbd_1	numeric	62%
msba_dbd	numeric	62%
msba_nbd_1	numeric	62%
msba_nbd_2	numeric	62%
msba_nbd_mean	numeric	62%
pgp_dbd_mean	numeric	62%
pgp_nbd_mean	numeric	62%
pgp_ratio	numeric	62%
msba_ratio	numeric	62%
msbaNBD_to_pgpDBD_ratio	numeric	62%
msbaNBD_pgpNBD_ratio	numeric	62%
bcrpNBD_to_pgp_DBD_ratio	numeric	62%
bcrpNBD_pgpNBD_ratio	numeric	62%
p450_mean	numeric	62%
notes	factor	0%
TARGET	factor	0%
molport.number	factor	0%
Works.on.Pgp	logical	0%
Actual.Toxicity..1.is.lethal...0.is.no.effect.	numeric	5%
Pgp.15.micromolar.efficacy..higher.number.the.better.	numeric	5%

X10.micromolar.bcrp.fold.inhibition..higher.number.is.better.	numeric	56%
X10.micromolar.P450.Inhibiton..closer.to.zero.is.better.	numeric	26%
MW	numeric	60%
xlogP	numeric	60%
Apolar.Desolvation..kcal.mol	numeric	60%
polar.desolvation..kcal.mol.	numeric	60%
H.bond.donors	integer	60%
H.bond.Acceptors	integer	60%
tPSA	integer	60%
Net.Charge	integer	60%
Rotatable.Bonds	integer	60%
SMILES	factor	0%
MOLECULEID	factor	0%
logPow..predicted.by.ochem.eu.model.4.in.Log.unit.	numeric	62%
Aqueous.Solubility..predicted.by.ochem.eu.model.4.in.log.mol.L..	numeric	62%
AMES..predicted.by.ochem.eu.model.1.	factor	0%
Numeric.prediction.for.AMES..predicted.by.ochem.eu.model.1.	numeric	62%
CYP450.3A4.modulation..predicted.by.ochem.eu.model.163.	factor	0%
Numeric.prediction.for.CYP450.modulation..predicted.by.ochem.eu.model.163.	numeric	62%
CYP450.modulation..predicted.by.ochem.eu.model.162.	factor	0%
Numeric.prediction.for.CYP450.modulation..predicted.by.ochem.eu.model.162.	numeric	62%
CYP450.modulation..predicted.by.ochem.eu.model.161.	factor	0%
Numeric.prediction.for.CYP450.modulation..predicted.by.ochem.eu.model.161.	numeric	62%
CYP450.modulation..predicted.by.ochem.eu.model.160.	factor	0%
Numeric.prediction.for.CYP450.modulation..predicted.by.ochem.eu.model.160.	numeric	62%
CYP450.modulation..predicted.by.ochem.eu.model.159.	factor	0%
Numeric.prediction.for.CYP450.modulation..predicted.by.ochem.eu.model.159.	numeric	62%
Aqueous.Solubility..predicted.by.ochem.eu.model.511.in.log.mol.L..	numeric	62%
X		100%
X.1		100%
X.2		100%
X.3		100%

X.4		100%
X.5		100%
X.6	numeric	86%
X.7	numeric	86%
X.8	numeric	86%
X.9	numeric	86%
X.10	numeric	86%
X.11	numeric	86%
X.12	numeric	86%
X.13	numeric	86%
X.14	numeric	86%
X.15	numeric	86%
X.16	numeric	86%
X.17	numeric	86%
X.18	numeric	86%
X.19	numeric	86%
X.20	numeric	86%
X.21	numeric	86%
X.22	numeric	86%
X.23	numeric	86%
X.24	numeric	86%
X.25	numeric	86%
X.26	numeric	86%
X.27	numeric	86%
X.28	numeric	86%
X.29	numeric	86%