Universidad Católica "Nuestra Señora de la Asunción"
Faculty of Science and Technology

Final Project submitted for the degree of Informatics Engineering

A Machine Learning Approach for the Identification of a Treatment against Chagas Disease

Student:

Rubén Jiménez

Supervisors:

Prof. Alberto Paccanaro
Prof. Luca Cernuzzi

October, 2017
Asunción, Paraguay

Acknowledgements

Foremost, I would like to thank my main supervisor Prof. Alberto Paccanaro for introducing me into the captivating field of bioinformatics. For his patient guidance and enthusiasm during this project. I would also like to thank Prof. Luca Cernuzzi, my co-supervisor, for his advice and useful critiques of this research work.

Special thanks to my co-worker Víctor Yubero for sharing a part of his research that is very valuable to this project. I would also like to acknowledge the Paccanaro lab members, Juan Cáceres, Mateo Torres, Diego Galeano, and former lab member, Dr. Horacio Caniza for their active collaboration in this work. I also wish to extend my thanks to the biologists from the CEDIC for their insight and knowledge shared with this project.

This research is part of the Consejo Nacional de Ciencia y Tecnología (CONACYT) grant 14-INV-088 "Identificación de cócteles de drogas para el tratamiento de la enfermedad de Chagas". This work was also supported by the Engineering and Physical Sciences Research Council (EPSRC) funding to Prof. Alberto Paccanaro, Computer Science Dept., Royal Holloway, University of London.

For my personal gratitude I would like to first thank God for giving me strength and guiding me all these years of study. I wish to extend my gratitude to my friends and colleagues that were right beside me during these challenging, but fun, years of university. I am particularly grateful to my fiancée Gabriela for giving me her unconditional love and support and always believing in me. I am very grateful for the warmth and sincere support of my family, I specially thank my sister Viole for her curiosity and interest in my research from her medical point of view; my brother Juanjo for all the discussions about this project that were of great help. Finally I wish to dedicate this thesis to my parents because I would not be who I am without their immense love, confidence and endless support.

Contents

Chapter 1

Introduction

Chagas disease is caused by the protozoan parasite *Trypanosoma cruzi* (*T. cruzi*). About 6 to 7 million people are infected with the parasite [1], and around 40 million are at risk of infection [2]. It is endemic throughout Latin America and has spread to other countries, making it a worldwide issue [3]. In Paraguay alone there are around 150.000 people suffering from this disease [4].

Insect vectors known as triatomine bugs are the primary mean of human infection. They typically live in the walls or roof cracks of poorly-constructed homes in rural or suburban areas [1]. After biting, they leave *T. cruzi* parasites (trypomastigotes) into excretion. These parasites are usually introduced into the bloodstream when the person instinctively smears the excretion into the bite wound or mucous membranes [5]. The disease can also be transmitted by transfusion of contaminated blood and congenitally from infected mothers to newborns [6].

The disease goes through two different phases: the acute, and the chronic phase. In the acute phase, symptoms are often absent or mild; they may include fever, headache, and enlarged lymph glands. Less than 50% cases develop characteristic symptoms: a skin lesion or a purplish swelling of the lids of one eye [1, 3]. In the chronic phase, parasites lodge mainly in the heart and digestive muscles. This can lead to severe organ pathologies and ultimately death [5].

Two drugs are currently available for the acute phase: Nifurtimox and Benznidazole. This phase, however, often goes undiagnosed due to a lack of proper diagnostic methods, and the inherent absence of symptoms [7]. Clinically, the disease is most commonly encountered in the chronic phase [8]. Treatment for this phase is highly limited due to the low potency of these drugs against the parasites [6]. Furthermore, no drugs are known to be effective in the chronic phase [7].

Prevention of the disease has relied mainly on vector control and prevention of transmission from non-vectorial mechanisms [9]. Control of Chagas disease also requires adequate antiparasitic treatment of chronically infected individuals [10]. New drugs with fewer side-effects and increased antitrypanosomal activity are urgently needed [3]. An impediment to solve this problem has been the lack of interest of pharmaceutical companies for the development of new antitrypanosomal agents [10].

Nevertheless, in the past years there has been increased interest in the disease; bringing major changes in the landscape of Chagas disease research and development. Institutions such as GlaxoSmithKline (GSK) and the Broad Institute are joining the effort to identify new therapies [11, 12, 13]. These institutions tackle this problematic by employing high-throughput screenings to test the activity of thousands of small molecules against *T. cruzi*.

High-throughput screenings (HTS) have been widely adopted by pharmaceutical companies over the past decades, with the goal of rapidly identifying potential drugs that affect specific molecular targets [14]. This type of assays are playing a major role in *in vitro* and *in vivo* clinical testing [15]. However, a limiting factor to this strategy is the expertise and resources required to develop each assay [16]. Additionally, lead molecules often have unfavourable side effects that are not discovered until later stages in the drug-discovery pipeline [14]. Technologies like HTS improve the rate and amount of information that can be collected about the effects of chemical compounds. This generates large-scale public datasets containing information on the structure and properties of small molecules, together with a score measuring their effectiveness on their targets.

The existence of these datasets allows the deployment of machine learning (ML) techniques for the prediction of potential drugs for a given disease [14]. Many machine learning methods are being proposed to improve and expedite the drug-discovery process. Some of these methods will be further discussed in the state-of-the-art section. Supervised machine learning are computational techniques that are able to learn complex relationships between noisy data (training set) and a given output (label), thus generating models that can be used for prediction. The vast amount of data generated from the screenings provides the ideal training data to develop a good predictive model. When applied to our problem, this translates into associating the small molecules, described by relevant molecular features, with their corresponding activity.

Employing models to establish a relationship between molecular descriptors and biological activity is known as quantitative structure-activity relationship (QSAR) analysis and plays a key role in virtual screenings [17]. These screenings are considered the best complement to HTS, because they are used to screen large libraries of small molecules reducing the number for further testing [18].

Drug repurposing or drug repositioning emerged as an attractive approach to find new uses of approved drugs. This strategy was already successfully used to find new therapies for neglected diseases, reducing the time, cost and risk of the drug-discovery process [19, 16]. Using ML predictive models to virtually screen drugs approved by the Food and Drugs Administration (FDA), may prove an effective method for repositioning or repurposing [16].

Another approach to asses the biological activity of new molecules is to measure the chemical similarity between molecules. The idea behind this method is that chemically similar molecules should lead to resembling biological activity. This approach has led to the finding of many valuable drugs in the past [20].

In this work, we focus on the problem of finding an effective treatment for Chagas disease at its chronic stage. We have developed two innovative approaches to select FDA approved drugs for their potential antitrypanosomal activity. In the first approach, we use the data from HTS assays against *T. cruzi* to train a machine learning model and virtually screen FDA approved drugs to predict their antitrypanosomal activity. In the second approach we identify the FDA approved drugs that are highly similar to the active molecules from HTS datasets at the chemical level. The objective is to find drugs with predicted activity against *T. cruzi* to guide a potential repurposing for Chagas disease. We also present an analysis based on the chemical substructures present in the molecules to derive a mathematical formulation that allows us to rank the set of molecules. The ranking is based on the contribution of each substructure to the experimental activity of the molecules.

To validate and evaluate the results from both approaches, we use a biologically moti-

vated analysis based on *T. cruzi* metabolic pathways. This method scores FDA approved drugs according to their potential for disrupting *T. cruzi* metabolic pathways by binding to the pathway enzymes. This analysis adds an extra layer of evidence and helps us to evaluate our results from a biological point of view. This analysis is a parallel research work which has been carried out within the same CONACYT project by my co-worker Víctor Yubero.

The present document is arranged in the following chapters:

- Chapter 2 introduces the necessary concepts needed to establish the background of the proposal.
- Chapter 3 includes the state-of-the-art of machine learning in the drug-discovery process, and specifically to Chagas disease drug-discovery.
- Chapter 4 describes our solution proposal with a description of the selected datasets and explains the implementation of the two approaches.
- Chapter 5 includes the results of our work and the analysis from a biological perspective.
- Chapter 6 ends the document with the conclusion and some ideas for future work that can be derived from this project.

1.1 Objectives

The main goal of this project is to develop a machine learning approach that can predict the biological activity of FDA approved drugs against *T. cruzi* to identify a potential treatment for Chagas disease.

We also propose the following specific objectives:

- Collect all available datasets of small molecules tested against *Trypanosoma cruzi*.
- Select and extract the molecular features to describe the dataset.
- Train different machine learning models and select the best performing one.
- Extract the molecular features of the FDA approved drugs that will be used for the prediction.
- Develop and apply a measure of chemical similarity between the FDA drugs and the active small molecules.
- Evaluate the set of predicted drugs from the biological point of view.

Chapter 2

Background

In this chapter by giving an overview of the drug-discovery process and introducing the definition of drug repurposing. Then we explain the concepts needed to implement a virtual screening device using high-throughput screening data and machine learning models. Finally, we show the concepts of our second approach and some other important biological concepts to understand the evaluation of our methods.

2.1 Drug-Discovery Process

The following steps are described in the FDA website [21]. They provide a general overview of the extensive process that drugs undergo to obtain approval. The scope of this work is enclosed in the first two steps of this process. Figure 2.1 shows the steps of the drug-discovery pipeline with the estimated time and cost of the process.

1. **Discovery and Development.** In this step, researchers discover potential new drugs through: new insights into a disease process that may lead to designing a product that stops or reverses the effects of the disease; many tests of molecules to find possible beneficial effects against a given disease.

   At this stage of the process, thousands, even millions of molecules may be potential candidates to develop new treatments. After initial testing and screenings, a reduced number of molecules may look promising and qualify for further study.

2. **Preclinical Research.** The promising molecules found in the previous step need to go through two types of preclinical tests. *In vitro* tests are usually conducted in a laboratory environment using test tubes, flasks or Petri dishes. Methods can be miniaturised and automated, yielding HTS methods for testing molecules in pharmacology or toxicology [22]. To further analyse the results of these tests and dismiss misleading results, *in vivo* testing is conducted in whole living organisms, usually lab animals.

   These studies must provide detailed information on dosing and toxicity levels. After preclinical research, the results are reviewed to decide whether the drugs should be tested in people.

3. **Clinical Research.** These trials are done in people to make sure the drugs are safe and effective. Only 10% of the drugs entering clinical trials get the FDA approval and reach the consumer market [18].

4. **FDA Review and Post-Market Safety Monitoring.** In this last step, the FDA examine all the submitted data to make a decision to approve or not to approve the drugs. All the approved drugs are monitored once the products are available for public use.

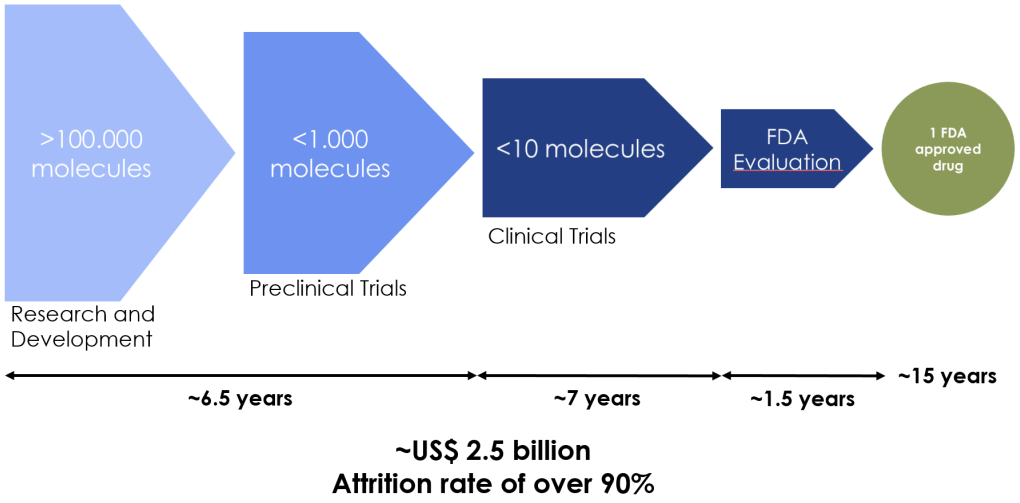


Figure 2.1: **Drug-discovery pipeline.**

The drug-discovery process is very expensive and requires a lot of time. It takes an average of 15 years and can cost up to US$2.5 billion, plus the attrition rate is over 90%, meaning that only 10% of the molecules entering the drug-discovery pipeline will be approved as a commercial drug [23]. The employment of computer-aided drug-discovery (CADD) techniques became essential for the preliminary stages of drug-discovery to expedite the process in a more cost-efficient way [24]. The field of CADD is rapidly improving with new methods and technologies being developed more frequently, revealing its huge potential in the drug-discovery pipeline [18].

2.2 Drug Repurposing

Drug repurposing or repositioning refers to the development of new indications for existing drugs [25]. During the past several years, there has been a surge of interest in this technique. Due to the big economic success of drug repurposing, pharmaceutical companies have recognised its advantages and the activity in this area has increased dramatically [26]. Table 2.1 shows successful examples of repurposing.

The development risk is reduced when either approved or discontinued drugs are candidates for repurposing [27]. The reason behind this is that these candidates have often passed through several stages of the drug-discovery process, and therefore have well-known safety profiles [26]. This allows the process of approval for the new indication to be expedited and reduced in costs and risks.

Effective drug repurposing requires that a known drug has a positive impact on a different disease. Its highest value resides in that its use for the novel indication surpasses the currently available therapeutic options for that condition [19]. This is a common

Drug Generic Name	Original Indication	New Indication	ref
Sildenafil	Angina	Male erectile dysfunction (Viagra)	[26]
Thalidomide	Sedation, nausea and insomnia	Cutaneous manifestations of moderate to severe erythema nodosum leprosum in leprosy and multiple myeloma	[26, 25]
Dapoxetine	Analgesia and depression	Premature ejaculation	[26]
Eflornithine	Anti-infective (originally for West African trypanosomiasis)	Reduction of unwanted facial hair in women	[26]
Finasteride	Prostate enlargement	Male pattern baldness	[25]
Ropinirole	Parkinson's disease	Restless leg syndrome	[25]

Table 2.1: **Examples of successfully repurposed drugs.** *Original indication* is what the drug was originally intended to treat. *New indication* is the new use of the drug.

scenario for neglected parasitic diseases, such as Chagas, where the available treatments are not effective.

To address this unmet medical needs, there have been numerous efforts to repurpose for parasitic diseases. A collection of these efforts are cited below:

- Amiodarone (antiarrhythmic) and Posaconazole (antifungal) tested against Chagas disease [28].
- Anti-cancer compounds studied *in vivo* against *T. cruzi* [29].
- Antiretroviral drugs tested against *Toxoplasma gondii* [30].
- Nitrofurazone (anti-infective agent) for the treatment of african sleeping sickness caused by *Trypanosoma Brucei* [31].

Historically, new indications for existing drugs were discovered fortuitously [27], like *Sildenafil* (Viagra). Today, there are new technologies that are aiding the development of systematic approaches to drug repurposing [25]. These technologies involve HTS and HCS[1] as well as database-driven bioinformatics techniques [19, 25].

2.3 High-Throughput Screenings

The workflow of an HTS can be conceptualized as a multi-stage funnel. In each stage, an experiment filters out uninteresting molecules and sends the rest to the next stage. The experiments are built with predefined thresholds to satisfy the goals of each project. These screenings are implemented during *in vitro* or *in vivo* assays (Step 2. of the drug-discovery

[1]High-content screenings (HCS) allow the automated acquisition and analysis of multiple cellular features simultaneously through image-based assays [32].

process) using robotic infrastructure to test thousands of small molecules against specific molecular targets [16].

HTS have become a major pillar in the drug-discovery process [15]. Pharmaceutical companies have used them extensively to identify and characterise bioactive molecules for a number of human diseases [33]. Specifically, in the area of infectious diseases, these screenings are being used to identify targets and mechanisms involved in the infective process of parasites [34, 35].

A few HTS assays have also been made against *T. cruzi* and are publicly available. These datasets are being applied to accelerate the discovery of new anti-parasitic drugs against Chagas disease, and could inspire new repurposing hypotheses. The most relevant articles presenting HTS assays against *T. cruzi* are presented below.

- *High Throughput Screening for Anti–Trypanosoma cruzi Drug Discovery* [32]. This article provides a literature review of HTS assays against *T. cruzi*. They analyse *in vitro* and *in vivo* phenotypic assays, as well as image based HTS and HCS assays in the hope of finding new anti–*T. cruzi* drugs.

- *Luminescence cell-based/microorganism primary HTS to identify inhibitors of Trypanosoma cruzi replication. PubChem BioAssay AID 1885* [36]. In this assay 303,224 molecules were screened, yielding 4,394 hits.

- *Image-Based High-Throughput Drug Screening Targeting the Intracellular Stage of Trypanosoma cruzi, the Agent of Chagas' Disease* [33]. In this work, the HTS has already been used to screen a small FDA approved drugs library (>900 compounds) where 55 hits were identified as potential candidates for repurposing.

- *New compound sets identified from high throughput phenotypic screenings against three kinetoplastid parasites: an open resource* [12]. In this article, GSK screened 1.8 million compounds against the three kinetoplastids most relevant to human disease: *Leishmania donovani*, *Trypanosoma cruzi* and *Trypanosoma brucei*. Consequently, three anti-kinetoplastid chemical boxes of around 200 compounds each were assembled.

2.4 Virtual Screenings

Virtual screenings are a result of combining computer-aided drug-discovery with HTS assays. They are used to screen large libraries of small molecules to detect a reduced number of lead molecules. These virtual screenings are considered the most popular complementary approach to HTS [18, 37]. The rich data generated from previous HTS assays is used to train machine learning models which are then deployed as virtual screening devices.

QSAR analysis is the name usually applied to the computational methods used to develop models employed as virtual screening devices. These methods correlate molecular structure (physicochemical properties) to some kind of *in vitro* or *in vivo* biological property — e.g. antitrypanosomal activity [17].

This method can work as a "virtual shortcut" to drug-discovery. By reducing the number of possible molecules that need experimental testing, virtual screenings can help in reducing the time and costs associated with this process. These devices are now routinely applied in campaigns to detect novel targets in the early stages of the drug-discovery pipeline (Step 1. of the drug-discovery process) [38].

Different machine learning techniques can be used to model the relationship between molecular descriptors and the experimental biological activity of molecules. Although virtual screenings are mostly applied to libraries of small molecules, in this project we aim to construct a QSAR model to virtually screen FDA approved drugs to predict their biological activity against *T. cruzi*. This can be considered as a first step in the search of new treatments against Chagas disease.

2.5 Machine Learning

ML are statistical and computational techniques that are applied to learn complex relationships within the data and build predictive models [14]. The term "learning" refers to running a computer program to induce a model by using training data or past experience [39].

Instances of datasets used in ML are represented using a set of features. These features may be continuous, categorical or binary. If instances are associated with known labels (the correct outputs), then the learning is called "supervised". In contrast, in unsupervised learning, instances are unlabelled [40].

Supervised classification is an important problem that ML is well suited to address. In a classification problem, we have a set of elements separated into classes. For any instance of the set, a class is assigned according to the instance's features and some classification rules. In many real-life situations, the classification rules are unknown, and the only information is the set of labelled examples from past experience. Supervised classification paradigms are algorithms that induce the classification rules from the data [39]. Examples of these algorithms include decision trees, support vector machines, bayesian classifiers [40].

There has been an increasing number of ML methods proposed in the drug-discovery process [27, 41]. A major area of interest for ML is virtual screening. The availability of technologies like HTS that provide experimental data on the activity of small molecules, provides the ideal training data to develop ML models as virtual screening devices.

In order to induce the ML model, a set of features is required to relate the small molecules with the corresponding activity. Decades of chemistry research has led to the development of molecular descriptors that describe a range of properties of any conceivable molecule [42]. These descriptors are often topological descriptors and molecular fingerprints [37], and can be used as features in traditional machine learning models. The employment of ML techniques requires a minimum amount of training data in order to build a good predictive model, and this is often linked to HTS [18].

The resulting models can harness the potential of the training data to predict key property values of new molecules, prioritising them for follow-up testing [16]. Importantly, the assumption that similar molecules exhibit similar biological activity compared with dissimilar or less similar molecules is generally valid [41]. Therefore, these models have an increasingly important role in the early steps of drug-discovery [14, 37]. Another potential use of these models is to predict novel association between drugs and diseases, making this an important tool also in the drug repurposing pipeline [27].

Now we will briefly describe the theory behind the different supervised classification algorithms that we later use to implement our solution proposal.

2.5.1 Naïve Bayes

Predicting the biological activity of molecules from molecular features can be formulated as a classification problem, in which we classify the molecules into two classes (C_1 = active,

C_0 = inactive), given an n-dimensional vector of genomic features $\mathbf{x} = (x_1, x_2, \cdots, x_n)$.[2]

The Bayesian Decision Rule states that in order to minimize the average probability of a classification error, one must choose the class with the highest posterior probability, i.e., assign a feature vector $\mathbf{x}$ to the class C_k, such that: $C_k = \arg_{C_i}\max P(C_i|\mathbf{x})$, where C_i ranges over the set of classes. C_k is known as the maximum a posteriori (MAP) estimate. Using Bayes theorem, the posterior probability can be rewritten, as

$$P(C_k|\mathbf{x}) = \frac{p(\mathbf{x}|C_k) \cdot P(C_k)}{p(\mathbf{x})}.$$

Notice that the unconditional density $p(\mathbf{x})$ in the denominator does not depend on the class label; therefore, it does not affect the classification decision and can be omitted when computing $C_k = \arg_{C_i}\max P(C_i|\mathbf{x})$. Each of the priors, $P(C_i)$, can be easily estimated by computing the frequency with which each class occurs in the data. However, the evaluation of $p(\mathbf{x}|C_i)$ cannot generally be accomplished in the same way, especially if the number of features is high; it would require a set of data large enough to contain many instances for each possible combination of feature values, in order to obtain reliable estimates.

The idea behind Naive Bayes [43] is to make the simplifying assumption that the attribute values are conditionally independent, given the target values. The computation of each is thus made efficient by approximating it as a product of conditional probabilities

$$\begin{aligned} p(\mathbf{x}|C_i) &= p(x_1, x_2, \cdots, x_n|C_i) \\ &\approx p(x_1|C_i)p(x_2|C_i)\cdots p(x_n|C_i) \\ &= \prod_j p(x_j|C_i). \end{aligned} \tag{2.1}$$

Learning in Naive Bayes consists of estimating the various $P(C_i)$ and various $p(x_j|C_i)$ using equation 2.1, based on their frequencies over the training data. Clearly, the approximation in equation 2.1 becomes exact only in the event of stochastic independence between the various features, given the class.

In the case of stochastic independence, the covariance between two features is zero. Thus, the covariance between features is a measure of the deviation from the condition of stochastic independence and is indicative of the amount of approximation introduced by the Naive Bayes assumption.

Alternatively, the Bayesian Decision rule for two classes can be stated thusly:

$$\text{If } \frac{p(\mathbf{x}|C_1) \cdot P(C_1)}{p(\mathbf{x}|C_0) \cdot P(C_0)} > 1 \text{ then choose class } C_1 \tag{2.2}$$
$$\text{Otherwise choose class } C_0.$$

If we then introduce the Naive Bayes approximation, we can rewrite equation 2.2 as:

$$\frac{p(x_1|C_1) \cdot p(x_2|C_1) \cdots p(x_n|C_1) \cdot p(C_1)}{p(x_1|C_0) \cdot p(x_2|C_0) \cdots p(x_n|C_0) \cdot p(C_0)} > 1 \tag{2.3}$$

2.5.2 Support Vector Machines

SVMs [44] are supervised ML algorithms that facilitate compound classification. Typically, they are used for binary property or activity prediction — e.g. classify biological activity of molecules against a specific target.

[2] Bold letters denote vectors; $P(\cdot)$ denote probabilities; $p(\cdot)$ denote probability density functions.

First, the dataset is projected into a high-dimensional feature space where the molecules are represented as descriptor vectors. The hope is that the molecules become linearly separable. This projection is accomplished by using a kernel function, such as one of the following types: linear, polynomial, sigmoid and radial basis (RBF). There has been a lot of work that shows hoy RBF-based SVMs outperforms SVMs based on the other three kernels and, thus, it is widely used.

Once linearly separable, the two classes of molecules can be separated in this feature space by a hyperplane. In fact, there is an infinite number of such hyperplanes and SVM chooses the hyperplane that maximizes the margin between the two classes on the assumption that the larger the margin, the lower the error of the classifier when dealing with unknown data. The hyperplanes that define such margins are called "support hyperplanes", and the data points that lie on these hyperplanes are the "support vectors" [41].

2.5.3 Decision Trees

Decision trees give us a set of "rules" to provide a way to associate specific molecular features with biological activity. This approach has already been applied to problems such as predicting 'drug-likeness' and predicting specific biological activities.

A decision tree is normally described as a tree, with the root at the top connected by successive and directional links or branches to other nodes. Starting from the root, the tree splits from the single trunk into two or more branches. Each branch itself might further split into two or more branches. This continues until a leaf is reached, which is a node that is not further split. The split of a branch is referred to as an internal node of the tree. The root and leaves are also nodes. Each leaf node is assigned with a target property, whereas a non-leaf node (root or internal node) is assigned with a molecular descriptor that becomes a test condition which branches out into groups of differing characteristics based on the possible values. In decision trees, the links must be mutually distinct and exhaustive — i.e. one and only one link will be followed. Figure 2.2 shows an example of this approach.

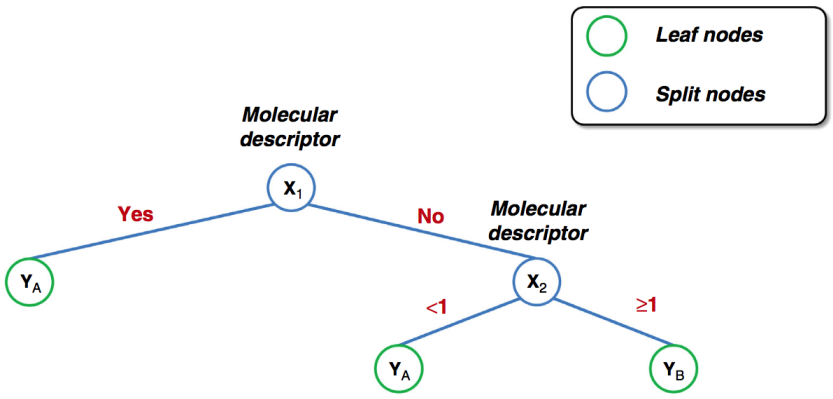


Figure 2.2: **Decision tree [41].** In the example, molecules with target properties Y_A and Y_B are classified based on two descriptors, X_1 and X_2

The tree construction process focuses on selecting the best test conditions to expand the extremities of the tree. The quality of the test conditions is usually determined by the "purity" of a split, which is often computed as the weighted average of the purity values of each branch, where the weights are determined by the fraction of examples that follow that branch. The metrics (e.g., information gain) used to select the best test generally

prefer test conditions that result in a balanced tree, where purity is increased for most of the examples, over test conditions that yield high purity for a relatively small subset of the data but low purity for the rest. Entropy, information-gain ratio, or Gini diversity index can be used as measure for the best classification.

Decision trees models are simple to understand, interpret, and validate. Their predictions, however, are known to suffer from high variance. Often a small change in the data can result in a different series of splits, complicating the interpretation. This instability is the result of the hierarchical nature of the process: the effect of an error in the top split is disseminated down to all the splits below. The performance of a decision tree depends on the proper selection of a sequence of splitting attributes of the training set for different levels in the hierarchy. The splitting attributes need to be sorted according to decreasing order of merit or importance. It is essential that the most important attribute is used for splitting at the root node, and the next in the rank for the immediate descendants of the root, and so on [39, 41].

2.5.4 Ensemble Methods

A common process to limit high variance is pruning the tree using either model complexity parameters or cross-validation. Generally, a single decision tree does not provide a high-performance model. Ho [45] proposed the use of an ensemble of decision trees, each created using a subset of the total descriptor set to increase the variance of the predictions, which he called the "random decision forest".

Breiman [46] introduces the concept of bagging, as an acronym of Bootstrap AGGregatING. The idea behind bagging is simple, that the ensemble is made of classifiers built using a unique base classifier on bootstrap replicates of the training set. The classifier outputs are combined by the *majority vote*[3]. To make use of the variations in the training set, the base classifier should be unstable, that is, small changes in the training set should lead to large changes in the classifier output. One of the most unstable classifiers are decision trees, as we have mentioned in the previous section.

Ensemble techniques are better predictors than an individual constituent learner and benefit from variability in the ensemble members, and so they take advantage of the variance of decision trees [39, 41]. In the next two sections we will go through two famous ensemble methods we use in this project, AdaBoost and Random Forest.

2.5.5 AdaBoost

A traditional way to combine the same base classifier is the AdaBoost algorithm [47]. The term comes from ADAptive BOOSTing. The general idea is to develop the classifier team incrementally, adding one classifier at a time. The classifier that joins the ensemble at one step is trained in a dataset selectively sampled from the initial training data set. The sampling distribution begins uniformly, and progresses towards increasing the likelihood of including "difficult" data points. Thus the distribution is updated at each step, increasing the likelihood of the objects misclassified at the previous step [39].

[3]The unseen instance will be classified as the class that obtains more votes from the different base classifiers whose output labels are fused.

2.5.6 Random Forest

The modern adaptation of the random decision forest, is the random forest algorithm developed by Breiman [48]. It introduced bagging and subset selection at each node of the decision tree to increase further prediction variance. Random forest is an ensemble classifier comprising many decision trees. Its architecture can be observed in Figure 2.3. Many classification trees are grown during training. A training set is created for each tree by random sampling with replacement from the original data set. During the construction of each tree, approximately one third of the cases are left out of the selection and this becomes the out-of-bag cases that are used as a test set. The classification performance of the test set is evaluated based on the out-of-bag error rates. Features will not be deleted based on one decision or one tree, but many trees will decide and confirm elimination of features.

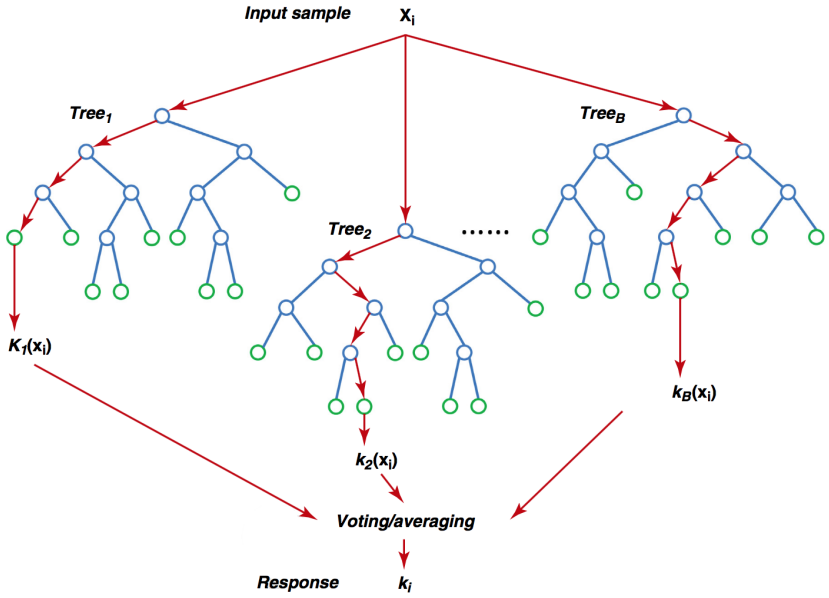


Figure 2.3: **Random Forest [41].** A general architecture of a random forest. Tree structures indicate yes/no rules at each branching. Individual predictions from all trees are collected and combined as a single ensemble prediction by voting for classification.

Another positive characteristic of random forest is that it is applicable to high-dimensional data with a low number of observations, a large amount of noise, and high correlated variables. Moreover, random forest is less prone to overfitting and can handle the problem of imbalanced classes. Random forest models have been proved to further increase the virtual screening performance of individual decision trees. Furthermore, random forest have been found to improve the prediction of QSAR analysis [49]. These properties include relatively high accuracy of prediction, built-in descriptor selection, and a method for assessing the importance of each descriptor to the model [41].

2.6 Chemical Similarity

Another approach that plays a major role in QSAR analysis is the measurement of chemical similarity. The goal of finding similarity at a chemical level between molecules is to allow the assessment of biological activities. The basic notion underlying this method is the

similar property principle, which states that similar chemical structures should lead to similar biological activities. This fundamental assumption has been used in medicinal chemistry in the past and has led to many valuable drugs [20].

Many similarity measures have been developed to quantify the degree of resemblance between chemical structures. These measures require a representation of the molecules so that they can be easily compared and a similarity coefficient that provides the mathematical function to calculate the value based on their representation [50].

A common and generally accepted way to represent the structural descriptors of the molecules are 2D molecular fingerprints. These fingerprints encode the presence or absence of chemical substructures present within a molecule. These substructures are the chemically relevant molecular fragments that compose the molecule. When a bit is set to *on* (1), it represents the presence of a given molecular substructure. This type of representation allows a fast and simple search, making them ideal for handling large chemical databases [51].

For the computation of fingerprint-based similarity, the Tanimoto coefficient is the method of choice [52, 53]. The mathematical formula for the Tanimoto coefficient, also described as the Jaccard coefficient, between two molecules A and B is:

$$S_{A,B} = \frac{c}{(a+b-c)} \approx \frac{|A \cap B|}{|A \cup B|} \qquad 0 \leq S_{A,B} \leq 1 \tag{2.4}$$

where a is the number of *on* bits in molecule A, b is the number of *on* bits in molecule B and c is the number of bits that are *on* in both molecules. The more similar two molecules are, the closer the Tanimoto coefficient is to 1. Typically, a Tanimoto coefficient > 0.85 is considered highly similar and coefficient > 0.75 is considered similar for the purpose of detecting biological activity profiles [54].

2.7 Metabolic Pathways

A metabolic pathway is a coordinated sequence of chemical reactions by which cells transform initial source compounds into final target compounds. A pathway can be conceptualised as a machinery that needs to complete a series of steps in order to obtain the final product.

Enzymes are protein catalysts in charge of the chemical reactions occurring within the pathway. In each step of the pathway, enzymes convert source compounds (substrates) into target compounds (products) by attaching or detaching chemical groups from substrates [55].

To get the full grasp of these concepts, in Figure 2.4 we can observe the *glucose* to *xylitol* conversion metabolic pathway [56]. In this example, the big light blue box is a cell. The arrows indicate chemical reactions and the ellipses represent the enzymes in charge of the reaction. The small boxes at the beginning of arrows are the substrates and those at the end, the products after the enzymatic reactions. For example, *ribulose* is the substrate for the chemical reaction in charge of enzyme 6 (E6) and *arabitol* is the product of said reaction. *Glucose* is the initial source compound and *xylitol* is the final product.

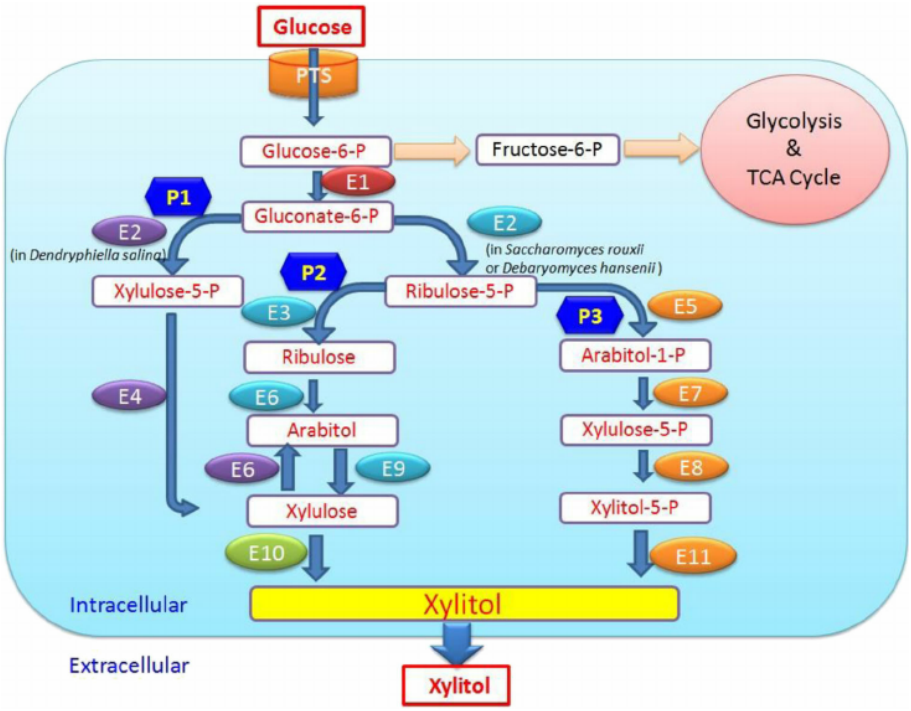


Figure 2.4: **Glucose to xylitol conversion pathway** [56].

Chapter 3

State of the Art

We now present a study of the literature to analyse the recent techniques employed to tackle our problematic and define the state-of-the-art. Literature searches were made in PubMed (https://www.ncbi.nlm.nih.gov/pubmed/) and in Google Scholar (https://scholar.google.com). In addition to the literature, important knowledge on the subject was shared in benefit to this research by a team of highly qualified biologists from the CEDIC (Centro para el Desarrollo de la Investigación Científica) http://www.cedicpy.com/.

3.1 Machine Learning in Drug-Discovery

Many authors have provided ML methods to increase the productivity of the drug-discovery process. Lavechia presents a review article [41] of the ML approaches in drug-discovery and an analysis of several recent publications, providing a detailed state-of-the-art in this field. In this section we present some of the most relevant examples of supervised classification algorithms employed in drug-discovery projects.

- **Decision trees**

  An implementation of decision trees, developed by Klekota and Roth, is used for the identification of substructures that discriminate activity from non activity within a given compound database [57]. Schneider et al. used decision trees for the classification of chemical compounds into drug and non drug and simultaneously for the purpose of deriving guidelines for the design of drug-like compounds [58]

- **Naïve Bayesian classifiers**

  Koutsoukas et al developed a Laplacian-modified Naïve Bayes classifier for in silico target prediction of bioactive molecules [59]. In a different study, Nigsch et al. also developed a Laplacian-modified Naïve Bayes classifier for ligand-target prediction [60].

- **SVM**

  Kinnings et al. in [61], describe how SVMs can be applied in developing non linear models of docking scoring functions using HTS data. They were able to increase the accuracy of virtual screening of direct inhibitors of *Mycobacterium tuberculosis* by using SVMs. Heikamp and Bajorath used weighted support vector machine linear combinations to predict compounds with closely related activity profiles [62]. Furthermore, in a drug repositioning attempt, Napolitano et al present a ML approach

(SVM) through data integration to predict drug therapeutic classes. The novelty of this approach relies on the purposeful interpretation of classification mismatches as genuine repositioning opportunities [19].

3.2 Machine Learning in Chagas Disease Drug-Discovery

The applications of ML in Chagas disease drug-discovery are still very scarce. We will describe the publications that we found in this specific area and also explain the difference between them and our approach.

A series of articles were published by Castillo-Garit in collaboration with Celeste Vega, our lead biologist in the project [63, 64, 65]. In their work, they report using linear discriminant analysis (LDA) to model the relationship between bond-based linear molecular descriptors and the antitrypanosomal activity. Classification models were developed to perform a QSAR study that allowed the discrimination of new antitrypanosomal drug-like compounds. High accuracy has been reported in their models and several *in vitro* experiments were performed to corroborate the reliability of their classification. This method permits a good prediction of the biological property, increasing the likelihood of the discovery of new compounds minimizing effectively the use of resources.

The main distinction with our approach is in the application of the QSAR model. Their aim is to predict new drug-like compounds to start developing a new drug against *T. cruzi*. On the other hand our aim is to identify FDA approved drugs with potential antitrypanosomal activity. By selecting drugs that are already approved, we know that the safety profiles are well studied and they are relatively safe for human consumption. We also propose the application of different ML algorithms to develop a classifier for the QSAR model.

An approach that is related to ours was developed by Ekins et al. [66] who utilised HTS data from the Broad Institute to develop Bayesian ML models to predict anti-parasitic activity against *T. cruzi in vitro*. They used these models to virtually screen 7200 molecules, including FDA approved drugs, to identify lead molecules with potential activity that may have not been tested yet. Molecules with the highest Bayesian score were selected and purchased. Less than 100 molecules were then tested *in vitro*, 11 of them were found actives and 5 qualified for an *in vivo* efficacy model testing. One of the molecules was found to have promising *in vivo* activity in the mouse model of Chagas disease. Our approach applies a different ML algorithm to develop a model that only classifies FDA approved drugs. We also add an evaluation of the results from a biological perspective. This evaluation is used to prioritise drugs for experimental testing.

3.3 Concluding Remarks

With the exponential growth of datasets, the application of ML methods in the drug-discovery process will continue to grow and be of more importance and influence. The experimental results are encouraging information, for our work, on how a ML model can be used as a virtual screening device to detect lead molecules. By following this workflow, the lead molecules can rapidly progress all the way to *in vivo* animal models and may lead to clinical studies in a shorter time scale.

After this literature study, we believe that there is still room left for improvement and a lot of work to be done in the area of ML applied to Chagas disease drug-discovery.

There are rich published HTS datasets against *T. cruzi*, and many supervised classification algorithms with known success.

Chapter 4

Solution Proposal

Based on the existence of publicly available HTS datasets against *T. cruzi*, we formulate two different approaches. The first consists in developing a machine learning approach to identify drugs from the set of FDA approved drugs with possible antitrypanosomal activity that could potentially work as a treatment for Chagas disease by eradicating the *T. cruzi* during the chronic stage.

Our second approach consists in developing and applying a measure of similarity at the chemical level between the FDA approved drugs and the small molecules that had been found to be active against *T. cruzi* in the experimental screenings. The idea is to use this similarity measure to select drugs that are chemically similar to these active small molecules. From this approach, we also develop an analysis based on the chemical substructures present in the molecules to derive a mathematical formulation that allows us to rank the set of molecules and prioritise the drugs that are chemically similar to the highest ranked molecules.

Finally, we analyse how the set of selected drugs can be justified and validated from the biological point of view. To do this, we relate the drugs to the *T. cruzi* metabolic pathways. This will allow us to prioritize our selection of FDA drugs that will be tested experimentally by a team of biologists at the CEDIC, as part of the CONACYT funded project.

4.1 The Broad Institute HTS Assay

A phenotypic HTS assay from the Broad Institute *T. cruzi* Inhibition Project [67] was selected. This types of assays are screenings based on the analysis of macroscopic effects on complex biological phenomena after treatment with libraries of small molecules. These phenomena include cell viability or genetic/metabolic responses and do not contemplate any *a priori* molecular target [15]. This particular assay detects the production of b-galactosidase by a genetically modified *T. cruzi* co-cultured in NIH3T3 cell lines. The method for the assay was a luminescent reporter system where a hit is detected if a molecule suppresses significantly the luminescence. This is the dose confirmation from a previous and more general assay detailed in [36]. The results contains **4063** molecules where **1852** (45%) were experimentally found to be actives against *T. cruzi*.

4.2 The GlaxoSmithKline HTS Assay

Using whole-cell phenotypic assays, GlaxoSmithKline performed an HTS where 1.8 million compounds were screened against the three kinetoplastids most relevant to human disease, i.e. *Leishmania donovani*, *T. cruzi* and *Trypanosoma brucei* [12]. Consequently, three anti-kinetoplastid chemical boxes of approximately 200 compounds each were assembled. Functional analyses of these compounds suggest a wide array of potential modes of action against the three parasites. This was the first published parallel HTS of a pharma compound collection against kinetoplastids. The boxes of small molecules are published as an open resource for future lead discovery programs. We have selected the *T. cruzi* chemical box with 222 small molecules exclusively for our chemical similarity approach that only cares about active molecules.

4.3 DrugBank

To mine the drug information we need for both approaches, we have selected the DrugBank database [68]. This is an unique bioinformatics and cheminformatics resource that combines detailed drug data with comprehensive drug target information. The database contains over 9000 drug entries including around 2000 FDA approved drugs.

4.4 Machine Learning Approach

Based on the selected HTS dataset, we formulate the problem as a classification problem, where our goal is to predict which FDA approved drugs are active against *T. cruzi*. Our fundamental assumption is that the activity of drugs and small molecules can be predicted from their chemical structure features. Therefore, we are using a machine learning algorithm to train a model that could associate the molecular features of the small molecules with their experimentally verified activity. We then apply this trained system to virtually screen FDA approved drugs.

A scheme of our approach is summarised in Figure 4.1. The specific steps involved in this process are the following (capital letters below refer to arrows in figure 4.1):

1. Formulate the problem as a classification problem.
2. Extract the features for our training set **(A)**.
3. Select a supervised classification algorithm.
4. Train the ML model **(B)**.
5. Test it using 10 fold cross-validation.
6. Extract the same features for the full set of FDA approved drugs **(C)**.
7. Use the model to virtually screen FDA approved drugs **(D)**.

We iterate through steps 3 to 5 refining our training parameters to select the best model for the prediction. Once the model is obtained, we use it to virtually screen the full set of FDA approved drugs to obtain their predicted activity against *T. cruzi*.

The success of ML models as predictors of biological activity greatly depends on the molecular descriptors selected to describe the small molecules from our dataset. We now

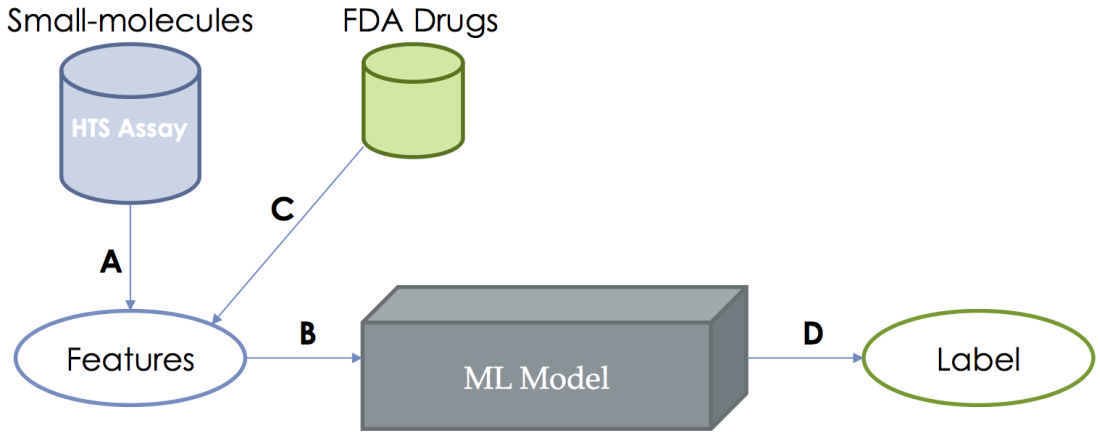


Figure 4.1: **The machine learning approach.** **A**. Extraction of the features for our dataset. **B**. Train the ML model. **C**. Extract the same features for the set of FDA approved drugs. **D**. Use the model to virtually screen the FDA drugs and predict activity.

present the features we chose for our model, and the tools used to extract them. For the extraction of the molecular features to represent the FDA approved drugs, the same procedure explained here was followed.

- **Function class fingerprints (FCFP_6)** are representations of chemical structures designed to capture molecular features relevant to biological activity. They represent the substructures that compose the molecule as a feature. The FCFP is able to represent up to 2^{32} molecular features. But for any given molecule only a subset of those features will be present. The appended number, next to the underscore, is the effective diameter of the largest feature and is equal to twice the number of iterations performed; for example, if three iterations are performed, the largest possible fragment will have a width of 6 bonds, and the fingerprint name will end in "_6".

  For our specific dataset, after calculating the FCFP for all the small molecules, we have identified over 26,000 molecular features present in the dataset. Each of those molecular features would translate to a feature for our ML model, but it is not reasonable to fit a model with over 26,000 features to only 4,000 data points.

  However, there is another way to represent the FCFP: a binary fingerprint of fixed length — e.g. 1024 bits. This fingerprint can be generated by hashing many molecular features into a smaller space. There is evidence that only a small amount of information is lost by this "folding" operation. But as the collision rate (two different substructures being represented by the same bit) is higher, the quality of the results will suffer [69]. For our models we have decided to use a fixed length of 128 bits for the FCFP, this way we can have a good amount of molecular features and still a reasonable number of features considering our dataset.

- **LogP** is the partition coefficient, and it defines the ratio of solubility of a compound in two immiscible solvents (usually water and octanol). It is an important absorption, distribution, metabolism and excretion (ADMET) feature for the molecule [70].

We have decided to use two representations of this measure, since they capture different aspects of the feature. The first one is the Atomic logP (ALogP) or Crippen logP because it was first proposed by Crippen et al. in [71]. AlogP considers that each atom contributes to the logP, and the molecule's final value is purely additive.

The other measure is the enhanced atomic or hybrid logP, known as XlogP. This is a modification of the AlogP system. It takes the value of each atom type, as well as a contribution from its neighbours and correction factors which help sidestep known deviances in purely atomistic methods. This method is very fast, and many free software tools use it.

- **Molecular weight** is the mass of a molecule. It is calculated as the sum of the atomic weight values of the atoms in a molecule. Its the mass of one mole of a substance, expressed in grams/mole [72].

- **Rotatable Bond Count** is the number of bonds that allow free rotation around themselves. It is considered an important predictor of good oral bioavailability. Is a topological parameter which measures the molecular flexibility [73].

- **Number of rings and aromatic rings** are useful characteristics of defining drug-likeness compounds [72].

- **Number of hydrogen bond acceptors/donors** They are considered important predictors of oral activity of drugs in humans [73].

- **Topological Polar Surface Area (TPSA)** is the surface summation over all polar atoms. It is related to the drug ability to permeate cells membranes (caco-2 permeability). It has also been shown to be a very good descriptor of absortion [72].

Features 1, 2 and 5 were calculated using the RDKit: "Open-Source Cheminformatics Software" (`http://www.rdkit.org`) from the Python Application Programming Interface (API). Features 3, 4, 6 and 7 were extracted from PubChem [74] using the *Entrez* package from the biopython module. The AlogP was extracted from ChemSpider (`www.chemspider.com`) querying the web API using the Python wrapper *ChemSpiPy*.

The extraction and processing of the molecular features were structured in CSV and ARFF files with 137 features in total, divided into 128 FCFP and 9 other molecular descriptors. With the training set ready, the next step is to train the ML model using different supervised classification algorithms and find the best one that can properly fit the 137 features into our 4063 data points from the Broad Institute HTS.

The models were trained and tested using Weka 3.8 [75]. In figure 4.2 we can observe the *KnowledgeFlow* of our approach. The first operator is the CSVLoader where we select as the input our processed training set with the 137 features. Next, the Class Assigner and Value Picker where we select the Label to be the class to predict and in the Picker we choose the value "Active" to later plot the receiver operating characteristics (ROC) curve. Our next operator is used to implement 10-fold cross validation. We can also observe the operators corresponding to the ML algorithms implemented. The last two operators, Performance Evaluator and Chart are to evaluate the models and visualise performance charts like the ROC curve.

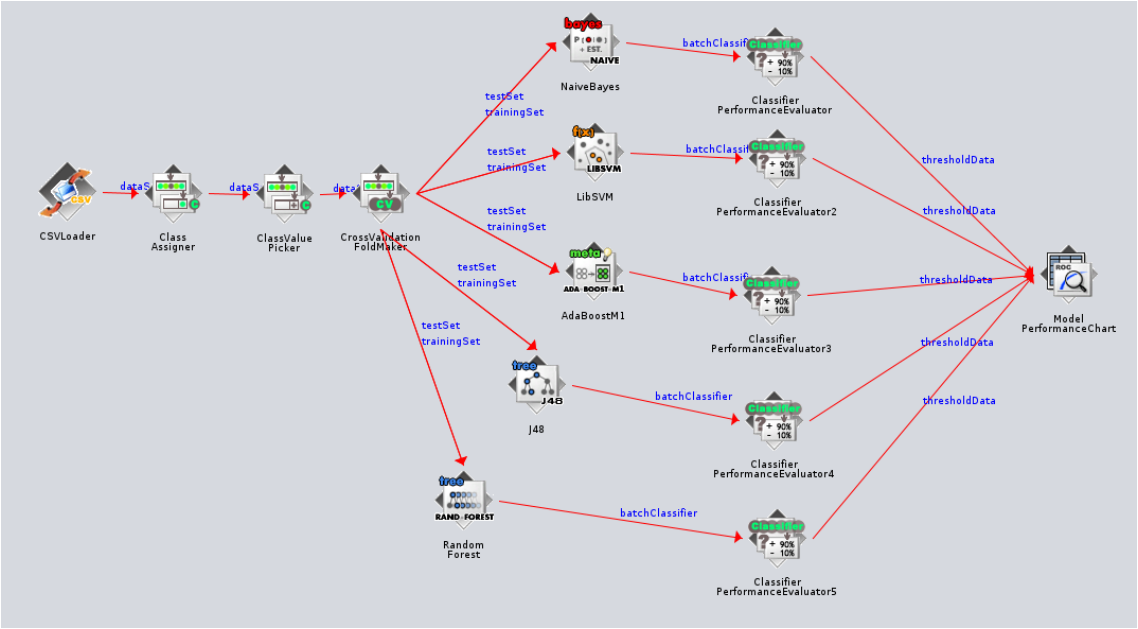


Figure 4.2: **Weka KnowledgeFlow of the ML approach implemented.**

4.5 Chemical Similarity Approach

For our second approach we want to quantify the degree of resemblance at the chemical level between active molecules, from our HTS datasets, and FDA approved drugs. Identifying a high chemical similarity between an active molecule and a drug could translate into similar biological activity.

To measure the similarity at the chemical level we use the Tanimoto coefficient. Our approach is summarised in Figure 4.3. First, we extract the active molecules from both HTS datasets. Then we calculate the Tanimoto coefficient between them and the FDA approved drugs. The results greater or equal than 0,75 are selected as analogues. We introduce the concept of "drug analogue" to denote an FDA approved drug that has a calculated Tanimoto coefficient greater or equal than 0,75 with an active molecule from the HTS datasets and therefore the two are considered highly similar at the chemical level. This approach was implemented using the RDKit: "Open-Source Cheminformatics Sotware" from the Python API.

4.6 Substructural Analysis

Data mining techniques are now widely used to identify relationships in large, multidimensional data sets in many areas and the analysis of HTS data is no exception. A key objective of such an analysis is the construction of models that enable relationships to be identified between the chemical structure and the observed activity. It is often more appropriate for HTS datasets to classify the molecules as "active" or "inactive" rather than using the numerical activity.

Our objective is to take advantage of the binary classification of HTS data and derive an analysis that quantifies the relevance of chemical substructures to the corresponding biological activity. This way we can create a ranking of substructures based on their contribution to the biological activity.

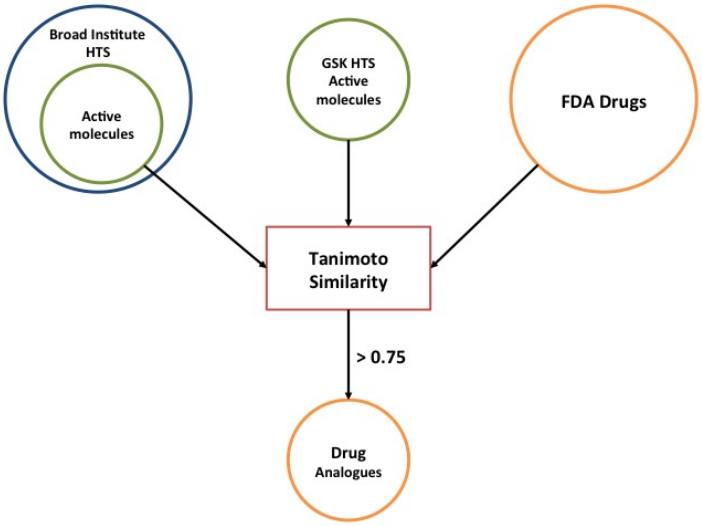


Figure 4.3: **The chemical similarity approach.** Drug analogues are identified by applying Tanimoto coefficient between active molecules from both HTS datasets and FDA approved drugs and selecting those with a score > 0.75.

The premise of substructural analysis (SSA) [76] is that each substructural fragment makes a constant contribution to the activity, independent of the other fragments in the molecule. The aim is to derive a weight for each substructural fragment that reflects its tendency to be in an active or an inactive molecule. The sum of the weights for all of the fragments contained within a molecule gives the score for the molecule. This enables the molecules to be ranked in decreasing probability of activity. Many different weighting schemes are possible. We define the weight of a fragment i according to:

$$w_i = \frac{act_i}{act_i + inact_i} \tag{4.1}$$

where act_i is the number of active molecules that contain the ith fragment and $inact_i$ is the number of inactive molecules that contain the ith fragment. This formula can be interpreted as the probability of the ith fragment to be included into an active molecule in the dataset. The fragments used in SSA often correspond to those present in structural keys of the type used for substructure searching such as the molecular fingerprints explained in the molecular features section above. The score for each molecule is the sum of the weight of the fragments or substructures contained in the molecule.

Our idea is to utilise the Broad Institute HTS dataset because it contains information about both active and inactive molecules. To extract the chemical substructures present in the dataset we use the FCFP from which we have identified over 26,000 distinct substructures. Note that for this approach we use the raw form of the FCFP with the full amount of substructures and not the fixed length bit string used for the ML approach. For this analysis we care about detecting which substructures contribute to the specific activity and the information lost in the hashing technique for the fingerprint could be very relevant.

4.7 Metabolic Pathway Analysis

For the evaluation and validation of the results from our approaches we use a parallel work from the same CONACYT project. This work is being developed by my co-worker Víctor Yubero as a part of his degree final project. It is based on an evolutionary approach to select drug cocktails for the treatment against Chagas disease. I will only provide here a brief and general overview of his method. The in-depth explanation of the techniques involved will be available in his thesis book.

The idea is to score drugs according to their potential to target *T. cruzi* enzymes, thus, increasing the likelihood of disrupting metabolic pathways. The role of an enzyme can be interpreted by the pathways it composes [77]. Figure 4.4 shows a basic explanation of our methodology. The first step (arrow A) is to obtain the set of protein sequences of *T. cruzi*. This was extracted from UniProt [78] which is a large database of protein sequences and associated detailed annotation. Then, we need to identify the metabolic pathways and enzymes present within *T. cruzi* (arrow B). Since experimentally validated metabolic pathways for *T. cruzi* are not publicly available, we inferred them computationally. We used the data from the Pathway Genome Data Base for *T. cruzi* to obtain the predicted metabolic pathways and enzymes. These data was published by Ekins et al in [66]. This research yielded 330 enzymes and 146 *T. cruzi* metabolic pathways.

To identify which drugs could potentially target *T. cruzi* pathways, we looked at the drug targets information from DrugBank. Drug targets come from a variety of organisms, and mainly from human. We can infer *T. cruzi* proteins, which will be affected by the drugs, by transferring information based on protein homology. Homologous proteins are thought to share important characteristics, such as their function and structure [79]. Furthermore, it can be assumed that homologous proteins will share their affinity to interact with a drug — i.e. to be targets of the same drug. High sequence similarity is an excellent indication of homology and, therefore, we computed the sequence similarity between the target proteins of every FDA approved drug and *T. cruzi* enzymes (arrow C). Basic Local Alignment Search Tool (BLAST) [80] was used to identify homologous proteins. The search identified 120 *T. cruzi* enzymes that are highly similar to drug targets. From this inference we were able to map 134 drugs from the full set of FDA approved drugs that have potential interaction with enzymes from 94 *T. cruzi* pathways (arrow D).

The mapped drugs we obtained with this analysis, may potentially affect *T. cruzi* by disrupting one or more metabolic pathways. We used this evolutionary approach to validate the results: we checked if any FDA drug which had been predicted as active against *T. cruzi* by our ML approach or by our chemical similarity approach targeted specific *T. cruzi* metabolic pathways. Predicting the same drug from differently motivated experiments is encouraging information about the potential antitrypanosomal activity of that drug.

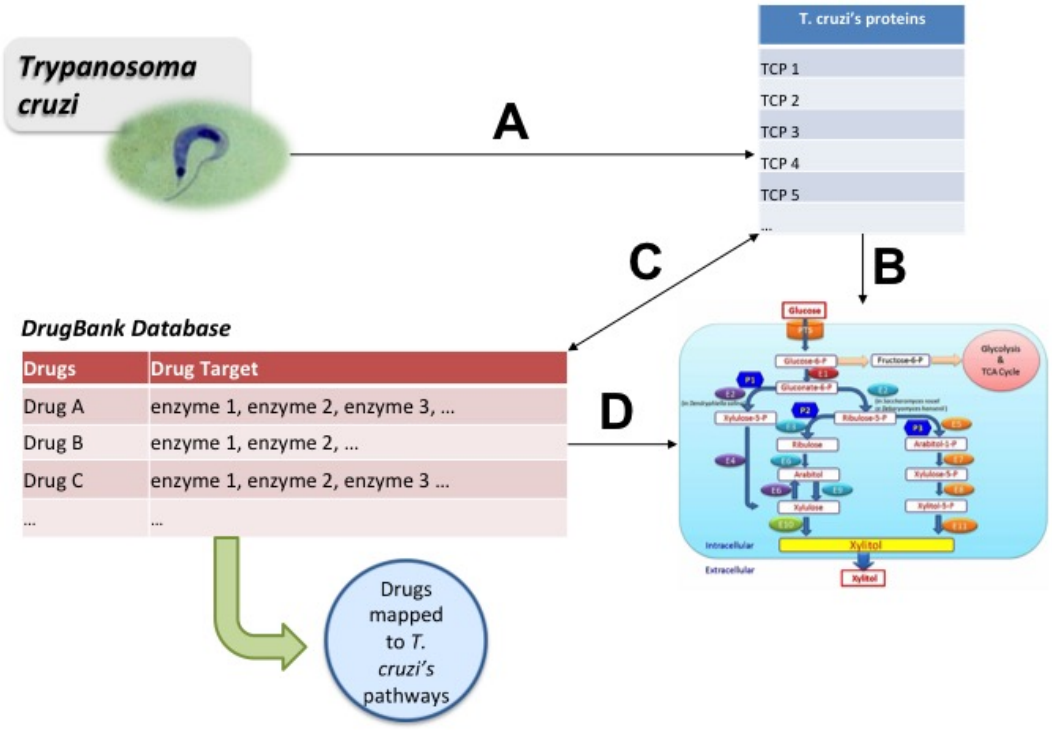


Figure 4.4: **Metabolic Pathway Analysis.** **A**. Breakdown *T. cruzi* into proteins. **B**. From the proteins, predict which pathways occur in *T. cruzi*. **C**. Find homologs between drug target proteins and *T. cruzi* proteins. **D**. Map resulting drugs to the pathways.

Chapter 5

Results

In this chapter we will show the analysis of our results and only the more relevant drugs from each approach. The full list of FDA approved drugs that resulted from each experiment will be uploaded and available in the project's website: http://www.dei.uc.edu.py/proyectos/proyectochagas/.

5.1 Machine Learning Approach

To assess the performance of each model we consider the accuracy of the classifier, the ROC curve and the area under the ROC curve (ROC AUC) as the metrics to guide our decision for the best ML model.

The accuracy of the classifier can be calculated as the correct classified instances from all predictions made. But this metric alone is not enough to evaluate and make a guided decision about the ML model to employ as a virtual screening device. The ROC curve is a graphical plot that shows the diagnostic ability of a binary classifier system as its discrimination threshold is varied. It is a well recognized metric used as an objective way to evaluate the ability of a classifier. To characterise both the ability of a virtual screening device to select active molecules and the ability to discard inactive ones, the ROC curve is well suited for this critical task.

The ROC curve applied to the analysis of a virtual screening experiment is a plot of the true positive rate (TPR) also called sensitivity, versus the false positive rate (FPR) also called 1-specificity, for all molecules in a ranked dataset. The TPR is defined as *TP/(TP+FN)* and the FPR is defined as *FP/(FP+TN)*. These values are derived typically from a confusion matrix like the one in Table 5.1.

A scoring function that would be able to perform perfect discrimination has a ROC curve that passes through the upper left corner of the plot, where the TPR is 1 (perfect sensitivity) and the FPR is 0 (perfect specificity). A random classification of the molecules would be represented by a diagonal rising from the origin to the upper right corner. Qualitatively, the closer the curve is to the upper left corner, the higher the overall accuracy of the test. The ROC AUC summarizes the overall performance of a virtual screening experiment [81, 82].

Principal component analysis (PCA) was also implemented to reduce the dimensionality of the feature space (137 features) and evaluate if it could lead to better performance. PCA is a mathematical algorithm that reduces the dimensionality of the data while retaining most of the variation in the data set. It accomplishes this reduction by identifying

		Prediction outcome		
		p′	**n′**	**total**
actual value	**p**	True Positive	False Negative	P
	n	False Positive	True Negative	N
	total	P′	N′	

Table 5.1: **Confusion Matrix** is a specific table that allows visualisation of the performance of a ML algorithm [83]. Each row of the matrix represents the instances in the actual class (p and n) while each column represents the instances in the predicted class (p′ and n′). The sum of each row equals the total number of instances of the actual class (P and N) and the sum of each column equals the total number of predictions for each class (P′ and N′).

directions, called principal components, along which the variation in the data is maximal. PCA identifies new variables, the principal components, which are linear combinations of the original variables. When the number of variables is too large compared to the number of samples (data points), PCA can reduce the dimensionality of variables without too much loss of information [84].

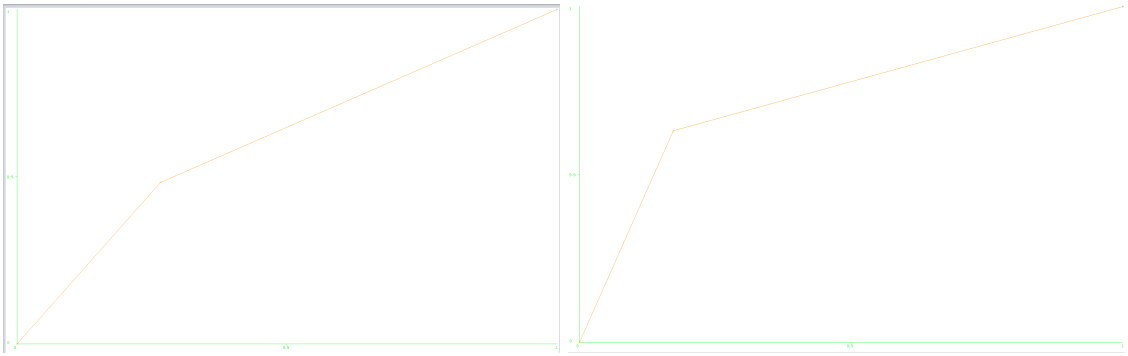
118 principal components are needed to retain 95% of the original variance. Although every algorithm was re-trained using the modified dataset with PCA, overall it did not improve any model. The only exception was the SVM model that improved dramatically after applying PCA. As mentioned before, this approach was implemented in Weka 3.8 and the *KnowledgeFlow* can be observed in Figure 4.2. We will now show the implementation for every algorithm and the performance metrics for every model.

- **Support Vector Machine**

  To implement this algorithm, the default parameters for the Weka libSVM were used. The (Gaussian) Radial Basis Function was the selected kernel. In Figure 5.1a we can observe the ROC curve for this model. After reducing the number of features with PCA, the performance of the algorithm experienced a notable boost. Figure 5.1b shows the ROC curve of the model with PCA. The accuracy climbed from 61,92% to 73,74% and the ROC AUC from 0,61 to 0,73.

- **Naïve Bayes**

  The default parameters were used to train the model. Figure 5.2 shows the ROC curve for this implementation. The accuracy of the model was 66,63% and the ROC AUC 0,693. After reducing the features with PCA, the performance of the model didn't suffer much alterations. The AUC remained the same and the accuracy improved in 0,5%. We can say that PCA barely had an impact on the performance of this model.



(a) **SVM model - AUC = 0,608.** (b) **SVM + PCA model - AUC = 0,73.**

Figure 5.1: **ROC curves for both SVM models.** Y-axis is the TPR (sensitivity) and the X-axis is the FPR (1-specificity).

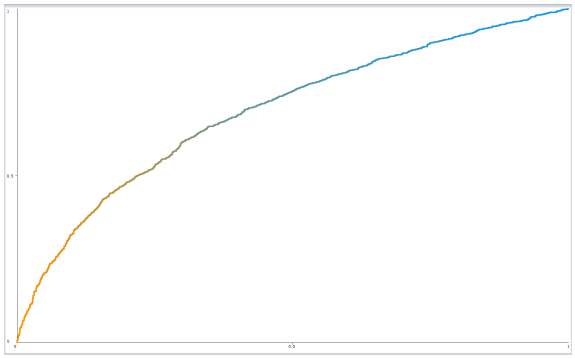


Figure 5.2: **ROC curve for the Naïve Bayes model - AUC = 0,693.** Y-axis is the TPR (sensitivity) and the X-axis is the FPR (1-specificity).

- **Decision Tree - J48**

  For our decision tree model we used J48 which is an open source Java implementation of the C4.5 algorithm [85] used in Weka. Figure 5.3a shows the ROC curve of the model. The accuracy of the model was 65,96% and the ROC AUC 0,66. After dimensionality reduction with PCA, both the accuracy and AUC diminished in more than 4%.

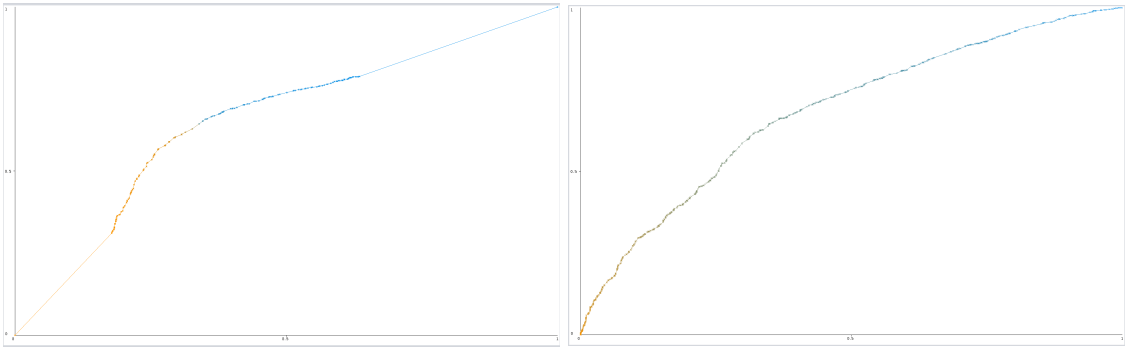
- **AdaBoost**

  As we described before, AdaBoost needs a different learning algorithm as base classifier to boost and improve their performance. We decided to use Decision Stump [86], which is a machine learning model consisting of a one-level decision tree. A decision stump makes a prediction based on the value of just a single input feature. This algorithm is selected often for its simplicity.

  The obtained accuracy of the model was 64,04% and the ROC AUC 0,68. After applying PCA, again, both the accuracy and AUC dropped by 2%. Figure 5.3b shows the ROC curve for this model.

- **Random Forest**

  The standard implementation of random forest in Weka was applied to obtain this model. Figure 5.4 shows the ROC curve for this model. The accuracy for the model



(a) **J48 model - AUC = 0,66.** (b) **AdaBoost model - AUC = 0,682.**

Figure 5.3: **ROC curves for J48 and AdaBoost models.** Y-axis is the TPR (sensitivity) and the X-axis is the FPR (1-specificity).

is the highest: **74,92%**; more than 10% better than any other algorithm. And the ROC AUC is **0,81** a number that proves how this model outperforms the others. After applying PCA to check if we could improve the performance, the result was the opposite. The accuracy dropped to 70,91% and the AUC to 0,76.

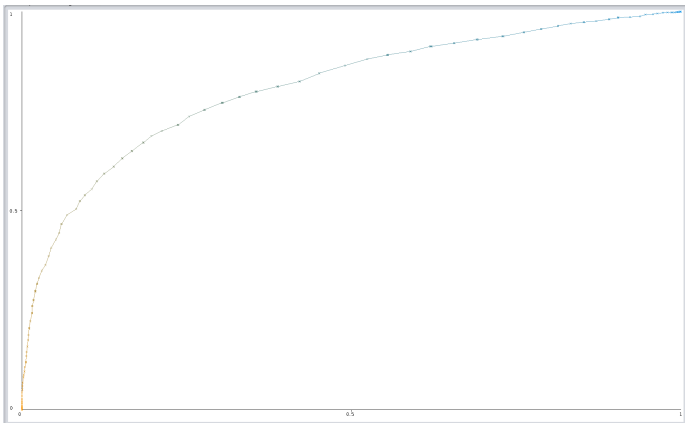


Figure 5.4: **ROC curve for the Random Forest model - AUC = 0,81.** Y-axis is the TPR (sensitivity) and the X-axis is the FPR (1-specificity).

ML Algorithm	**Accuracy**	**ROC-AUC**
Random Forest	74,92%	0,81
SVM*	73,74%	0,73
Naïve Bayes	66,63%	0,71
AdaBoost	64,04%	0,68
Decision Tree (J48)	65,96%	0,66

Table 5.2: **Comparison of the ML models.** **PCA was applied.*

The final part of this approach is to virtually screen FDA approved drugs and identify those with potential activity against *T. cruzi* (arrow **D** in Figure 4.1). After the analysis of the performance of the implemented models, detailed in Table 5.2, we chose the Random Forest model as our virtual screening device to classify the biological activity of the full set of FDA approved drugs. Weka 3.8 was used to obtain the classification using the previously trained model. A total of 63 FDA approved drugs were classified as actives. Table 5.3 presents a list of the top 10 drugs ranked by their classification score.

Ranking	Drug Generic Name	Classification Score
1	Telaprevir	0,95
2	Vemurafenib	0,95
3	Cephaloglycin	0,93
4	Thymol	0,82
5	Diethylstilbestrol	0,82
6	Choline C 11	0,78
7	Fluvastatin	0,78
8	Oxtriphylline	0,78
9	Hydroxychloroquine	0,77
10	Frovatriptan	0,77

Table 5.3: **Top 10 drugs classified as actives.** The score is equal to the probability of the drug of actually belonging to the predicted class.

Random forest is currently considered one of the best QSAR methods available in terms of accuracy of prediction according to the literature [87], so it is not strange that it is our best performing model. The performance of our model is comparable to the state-of-the-art classifier [66] in terms of accuracy and ROC AUC.

5.2 Chemical Similarity Approach

As described in Section 4.5, a drug analogue is selected only if the Tanimoto coefficient is greater or equal than 0,75. A total of 101 unique FDA approved drugs were identified to be analogues with active molecules from both HTS datasets, the Broad Institute and GlaxoSmithKline. In Table 5.4 we observe the pairs (drug, molecule) with the highest similarity based on the Tanimoto score.

It is important to emphasize that a Tanimoto coefficient of 1 does not necessarily mean that two compounds are identical. It only means that they have identical structural descriptors or identical on-bits in a binary fingerprint. However, there are cases in which a Tanimoto score of 1 identifies the same compound — e.g. Itraconazole.

We have obtained two lists of FDA approved drugs with potential antitrypanosomal activity. One list corresponds to the drugs classified as actives by our Random Forest ML model and the other list corresponds to the drug analogues. **Itraconazole, Monoben-**

Drug Generic Name	Molecule Name	Tanimoto Score
Paclitaxel	AC1Q5F8K	1,0
Chlorhexidine	MLS001304094	1,0
Itraconazole	Itraconazole	1,0
Verapamil	Verapamil Hydrochloride	1,0
Chlorhexidine	Chlorhexidine	1,0
Fusafungine	Enniatin B	1,0
Docetaxel	Docetaxel Intermediate	1,0
Cabazitaxel	Docetaxel Intermediate	0,99
Mitomycin	Porfiromycine	0,98
Tubocurarine	Trigilletine	0,96

Table 5.4: **Top 10 drug analogues according to their Tanimoto coefficient score with active molecules.**

zone and Bleomycin are the only drugs present in both lists. It is encouraging evidence to find the same drugs as a result from two differently motivated experiments. Moreover, we discovered that Itraconazole has already been clinically trialled against Chagas disease [88], obtaining good results in improving cardiac conditions but a higher efficacy and less side effects are needed before considering this drug as a viable treatment. Finding a drug with this much evidence of its antitrypanosomal activity really adds extra weight to the other results from both approaches.

5.3 Substructural Analysis

After applying substructural analysis on the Broad Institute HTS dataset, we obtained a ranking for the molecules. The top 10 ranked molecules can be observed in Table 5.5. Figure 5.5 shows the plot of the SSA score (Y-axis) versus the ordered molecules, from highest lo lowest SSA score, in the dataset (X-axis) divided in active (red) and inactive (blue). Analysing this plot, it is clear that if we trace a line somewhere around $Y = 40$ we can set a threshold that discriminates by the SSA score the active molecules from the inactive molecules.

Molecules with the highest score (SSA $>$ 40) contain fragments or substructures that are more relevant to activity and therefore, the drug analogues to these molecules have a higher chance to be effective and have activity against *T. cruzi*. We have identified a subset of drug analogues that are analogues to molecules with an SSA score $>$ 40, Table 5.6. We consider that these specific drugs are more likely to show antitrypanosomal activity based on the SSA ranking analysis. There are 9 drugs that are analogues with the two highest ranked molecule in the SSA ranking. Table 5.6 shows the three more chemically similar drugs (Tanimoto $>$ 0,9) that are analogues with the two best ranked molecules.

Our approach identifies Docetaxel as one of the most promising candidates for repur-

Ranking	Molecule Chemical Name	SSA-score
1	7-epi-Cephalomannine	70,62
2	2",3"-Dihydrocephlomannine	69,08
3	TG-IR	68,41
4	AC1Q5F8K	68,1
5	10-Deacetyl-7-epitaxol	67,1
6	10-Deacetyltaxol	67,1
7	Docetaxel Intermediate	66,65
8	DHPC1_000753	66,15
9	Mezerein	62,37
10	MLS000834504	62,3

Table 5.5: **Top 10 molecules ranked according to their SSA score.**

Drug Generic Name	Molecule Name	Tanimoto Score	SSA Score
Paclitaxel	7-epi-Cephalomannine	0,94	70,62
Docetaxel	7-epi-Cephalomannine	0,91	70,62
Cabazitaxel	2",3"-Dihydrocephlomannine	0,92	69,08

Table 5.6: **Top 3 drugs analogues with molecules with SSA score $>$ 40.**

posing and this same drug was already identified to have antitrypanosomal activity in the past [33]. This is promising evidence that our approach is well directed and is able to produce reasonable results.

5.4 Metabolic Pathway Analysis

We used our metabolic pathway analysis to evaluate and validate the results from both approaches. This will identify if any FDA approved drug classified as active with our ML model or any drug analogue target *T. cruzi* metabolic pathways.

- **From the FDA approved drugs classified as actives with the Random Forest model,** we have identified **3** drugs that target a total of **5** enzymes and **6** *T. cruzi* metabolic pathways:

  1. **Itraconazole** targets 4 pathways: *plant sterol biosynthesis*, *cholesterol biosynthesis I*, *cholesterol biosynthesis II* and *cholesterol biosynthesis III*.
  2. **Phosphatidylserine** targets the *phosphatidylethanolamine biosynthesis I* pathway.

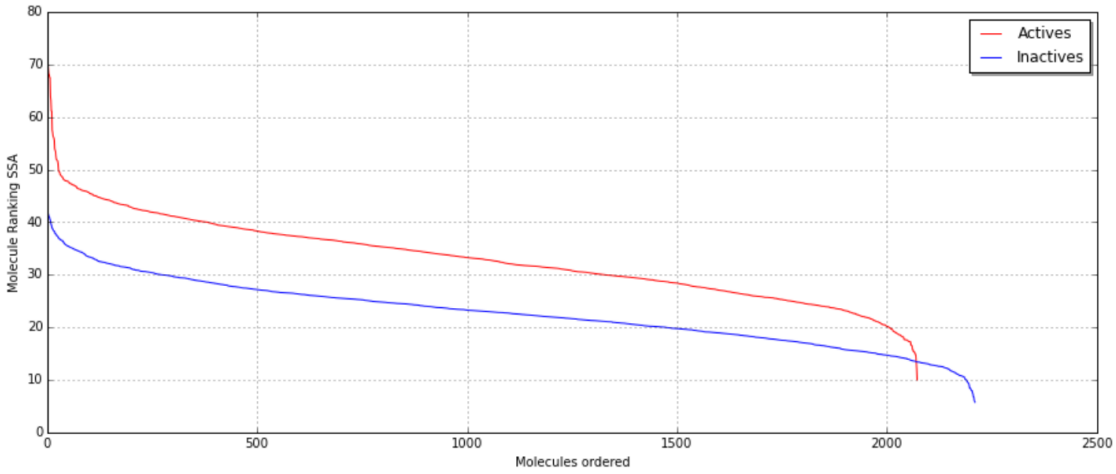


Figure 5.5: **Ranking of the molecules according to their SSA score.** Blue line represents the inactive molecules and the red line represents the active molecules.

3. **Ethionamide** targets the *ceramide de novo biosynthesis* pathway.

- **From the drug analogues,** we have identified **11** drugs that target a total of **30** enzymes and **34** *T. cruzi* metabolic pathways:

  1. **Adenosine monophosphate** targets 13 pathways: *pyruvate fermentation to acetate and alanine, fatty acid biosynthesis initiation II, octanoyl-ACP biosynthesis (mitochondria, yeast), crotonate fermentation (to acetate and cyclohexane carboxylate), pyruvate fermentation to acetate III, docosahexanoate biosynthesis II, fatty acids biosynthesis (yeast), acetate formation from acetyl-CoA II, docosahexanoate biosynthesis I, fatty acid activation, ethanol degradation II, acetate conversion to acetyl-CoA, and sucrose biosynthesis I (from photosynthesis).*
  2. **Flavin adenine dinucleotide** targets 8 pathways: *2-oxoisovalerate decarboxylation to isobutanoyl-CoA, 2-oxoglutarate decarboxylation to succinyl-CoA, aerobic respiration (cytochrome c), glycine cleavage, ethanol degradation I, pyruvate decarboxylation to acetyl CoA, ethanol degradation II and leucine degradation I.*
  3. **Itraconazole** targets 4 pathways: *plant sterol biosynthesis, cholesterol biosynthesis I, cholesterol biosynthesis II* and *cholesterol biosynthesis III.*
  4. **Cladribine** targets 5 pathways: *adenosine deoxyribonucleotides de novo biosynthesis, superpathway of adenosine nucleotides de novo biosynthesis I, pyrimidine deoxyribonucleotides de novo biosynthesis I, superpathway of guanosine nucleotides de novo biosynthesis I and guanosine deoxyribonucleotides de novo biosynthesis I.*
  5. **Nitrofural** targets 5 pathways: *2-oxoisovalerate decarboxylation to isobutanoyl-CoA, pyruvate decarboxylation to acetyl CoA, 2-oxoglutarate decarboxylation to succinyl-CoA, aspartate degradation II and glycine cleavage.*
  6. **Rifampicin** targets 4 pathways: *plant sterol biosynthesis, cholesterol biosynthesis I, cholesterol biosynthesis II (via 24,25-dihydrolanosterol) and cholesterol biosynthesis III (via desmosterol).*

7. **Ketoconazole** targets 4 pathways: *plant sterol biosynthesis, cholesterol biosynthesis I, cholesterol biosynthesis II (via 24,25-dihydrolanosterol) and cholesterol biosynthesis III (via desmosterol).*
8. **S-Adenosylmethionine** targets 4 pathways: *spermidine biosynthesis I, cysteine biosynthesis I, methionine degradation I (to homocysteine) and S-adenosyl-L-methionine biosynthesis.*
9. **Terconazole** targets 4 pathways: *plant sterol biosynthesis, cholesterol biosynthesis I, cholesterol biosynthesis II (via 24,25-dihydrolanosterol) and cholesterol biosynthesis III (via desmosterol).*
10. **Proguanil** targets the pyrimidine deoxyribonucleotides de novo biosynthesis I pathway.
11. **Adenosine triphosphate** targets 12 pathways: *pyruvate fermentation to acetate and alanine, fatty acid biosynthesis initiation II, octanoyl-ACP biosynthesis (mitochondria, yeast), crotonate fermentation (to acetate and cyclohexane carboxylate), pyruvate fermentation to acetate III, docosahexanoate biosynthesis II, fatty acids biosynthesis (yeast), acetate formation from acetyl-CoA II, docosahexanoate biosynthesis I, fatty acid activation, ethanol degradation II and acetate conversion to acetyl-CoA.*

It is very important that all of these drugs can be evaluated from the biological point of view by relating their potential activity to metabolic pathways, indicating their possible mechanism of action. This is no small accomplishment, since very little is known about *T. cruzi*. It is also important to mention that as a part of this work we can identify which enzymes are being targeted by these drugs. This information is very valuable to the biologists because it is a way to translate our pure computer scientist approach to terms they can easily understand and use for formulating hypothesis.

Itraconazole is the only drug that shows up in both approaches and that can also be evaluated and validated by the metabolic pathways analysis. As we mentioned before this is a drug that was already studied and trialled against Chagas disease. It is a very important fact that the only drug that came up in each of the different approaches that we developed reached advanced stages in the drug development process, demonstrating the potential of our other results.

Chapter 6

Conclusion

In this final degree project we have presented a machine learning approach to predict the biological activity of FDA approved drugs against *T. cruzi*. We believe that the proposed methodology will expand the state-of-art of machine learning in the Chagas disease drug-discovery pipeline. We have obtained similar performance results with the work presented in [66] but applied only to FDA approved drugs as a repurposing strategy. Also, the selection of Random Forest model based on its performance agrees with the literature that states that it is one of the best well known methods for predicting biological activity.

We have also developed a chemical similarity approach to identify FDA approved drugs that are analogues with the active molecules from two different HTS datasets. The Tanimoto coefficient was used to identify the drug analogues. Another contribution of this project is the substructural analysis performed on the Broad Institute dataset. This analysis produced a ranking of the molecules according to the contribution to biological activity of each fragment or substructure. Combining the chemical similarity approach and the substructural analysis allowed us to identify a subset of drug analogues with a higher chance of activity against *T. cruzi*. Although it is out of the scope of this work, SSA can also help to detect the substructures that contribute the most to activity, and they could be used for prioritising them for further studies.

A final contribution of this work is the biological evaluation provided by the metabolic pathway analysis. This evaluation allows us to map FDA approved drugs onto *T. cruzi* metabolic pathways. This validation is useful because it incorporates important information of how the drugs target *T. cruzi*.

Finding a subset of drugs that come up from differently motivated experiments is promising. The fact that among our results are drugs that already have been tested in the past against Chagas disease is encouraging evidence that our approaches are able to produce reasonable candidates for drug repurposing. Additionally, the majority of the drugs present in our results were never tested against *T. cruzi*, confirming the novelty of our approaches.

We believe that this work is a step forward in the battle against Chagas disease and that it will expand the landscape of drug-discovery for neglected parasitic diseases. As future work, the pipeline we presented in this work and the same ideas behind this research can be broadly applied to other neglected parasitic diseases — e.g. Leishmaniasis (caused by the protozoa *Leishmania*). The repurposing goal allows this pipeline to be adjusted to different organisms, as long as experimental HTS data is available.

As a part of this large CONACYT project, the biologists at the CEDIC are currently reviewing our results and are preparing to enter the preclinical stage where *in vitro* and *in*

vivo trials will be conducted on these drugs. We have already received positive feedback from them confirming their interest in the results we have produced.

Ideas for future work include the improvement of the ML models, further exploration and analysis of the feature space to produce more descriptive molecular features. Incorporating information of the 3D structure of the molecules and of drugs to predict if the drug will be able to bind to its target. This same idea could also be applied to refine the metabolic pathway analysis by improving the metric that is used to identify homologs.

Bibliography

[1] W. H. Organization, "Chagas disease fact sheet no. 340." Available at: http://www.who.int/mediacentre/factsheets/fs340/en/, March 2017. Accessed: 20.03.2017.

[2] C. J. Schofield, J. Jannin, and R. Salvatella, "The future of chagas disease control," *Trends in parasitology*, vol. 22, no. 12, pp. 583–588, 2006.

[3] A. Rassi and J. A. Marin-Neto, "Chagas disease," *The Lancet*, vol. 375, no. 9723, pp. 1388–1402, 2010.

[4] O. P. de la Salud, "Artículo sobre la enfermedad de chagas en paraguay." Available at: http://www.paho.org/par/index.php?option=com_content&view=article&id=677:enfermedad-chagas-calcula-50-000-nuevos-casos-ano-america-150-000-personas-infectadas-paraguay&Itemid=258. Accessed: 16.01.2017.

[5] I. Ribeiro, A. Sevcsik, *et al.*, "New, improved treatments for chagas disease: From the R&D pipeline to the patients," *Plos Neglect Trop D*, vol. 3, no. 7, p. e484, 2009.

[6] J. A. Urbina, "Specific chemotherapy of chagas disease: relevance, current limitations and new approaches," *Acta tropica*, vol. 115, no. 1, pp. 55–68, 2010.

[7] J. R. Coura and P. A. Viñas, "Chagas disease: a new worldwide challenge," *Nature*, vol. 465, no. n7301_supp, pp. S6–S7, 2010.

[8] S. R Wilkinson, C. Bot, J. M Kelly, and B. S Hall, "Trypanocidal activity of nitroaromatic prodrugs: current treatments and future perspectives," *Current topics in medicinal chemistry*, vol. 11, no. 16, pp. 2072–2084, 2011.

[9] J. R. Coura and J. C. P. Dias, "Epidemiology, control and surveillance of chagas disease: 100 years after its discovery," *Memórias do Instituto Oswaldo Cruz*, vol. 104, pp. 31–40, 2009.

[10] A. Rassi, J. P. Dias, and J. A. Marin-Neto, "Challenges and opportunities for primary, secondary, and tertiary prevention of chagas' disease," *Heart*, vol. 95, no. 7, pp. 524–534, 2009.

[11] E. Chatelain, "Chagas disease drug discovery: toward a new era," *Journal of biomolecular screening*, vol. 20, no. 1, pp. 22–35, 2015.

[12] I. Peña, M. P. Manzano, J. Cantizani, A. Kessler, J. Alonso-Padilla, A. I. Bardera, E. Alvarez, G. Colmenarejo, I. Cotillo, I. Roquero, *et al.*, "New compound sets identified from high throughput phenotypic screening against three kinetoplastid parasites: an open resource," *Scientific reports*, vol. 5, p. 8771, 2015.

[13] G. Andriani, A.-D. C. Chessler, G. Courtemanche, B. A. Burleigh, and A. Rodriguez, "Activity in vivo of anti-trypanosoma cruzi compounds selected from a high throughput screening," *PLoS Negl Trop Dis*, vol. 5, no. 8, p. e1298, 2011.

[14] R. F. Murphy, "An active role for machine learning in drug development," *Nat Chem Biol*, vol. 7, pp. 327–330, June 2011.

[15] I. Fraietta and F. Gasparri, "The development of high-content screening (hcs) technology and its importance to drug discovery," *Expert opinion on drug discovery*, vol. 11, no. 5, pp. 501–514, 2016.

[16] S. J. Swamidass, "Mining small-molecule screens to repurpose drugs," *Briefings in bioinformatics*, vol. 12, no. 4, pp. 327–335, 2011.

[17] D. A. Winkler, "The role of quantitative structure-activity relationships (qsar) in biomolecular discovery," *Briefings in bioinformatics*, vol. 3, no. 1, pp. 73–86, 2002.

[18] S. P. Leelananda and S. Lindert, "Computational methods in drug discovery," *Beilstein Journal of Organic Chemistry*, vol. 12, no. 1, pp. 2694–2718, 2016.

[19] F. Napolitano, Y. Zhao, V. M. Moreira, R. Tagliaferri, J. Kere, M. D'Amato, and D. Greco, "Drug repositioning: a machine-learning approach through data integration," *Journal of cheminformatics*, vol. 5, no. 1, p. 30, 2013.

[20] N. Nikolova and J. Jaworska, "Approaches to measure chemical similarity–a review," *Molecular Informatics*, vol. 22, no. 9-10, pp. 1006–1026, 2003.

[21] F. Food and D. Administration, "The drug development process." Available at: https://www.fda.gov/forpatients/approvals/drugs/, June 2015. Accessed: 05.04.2017.

[22] N. Quignot, J. Hamon, and F. Y. Bois, "Extrapolating in vitro results to predict human toxicity," *In Vitro Toxicology Systems*, pp. 531–550, 2014.

[23] J. Avorn, "The $2.6 billion pill—methodologic and policy considerations," *New England Journal of Medicine*, vol. 372, no. 20, pp. 1877–1879, 2015.

[24] S. J. Y. Macalino, V. Gosu, S. Hong, and S. Choi, "Role of computer-aided drug design in modern drug discovery," *Archives of pharmacal research*, vol. 38, no. 9, pp. 1686–1701, 2015.

[25] S. H. Sleigh and C. L. Barton, "Repurposing strategies for therapeutics," *Pharmaceutical Medicine*, vol. 24, no. 3, pp. 151–159, 2010.

[26] T. T. Ashburn and K. B. Thor, "Drug repositioning: identifying and developing new uses for existing drugs," *Nature reviews Drug discovery*, vol. 3, no. 8, pp. 673–683, 2004.

[27] J. Li, S. Zheng, B. Chen, A. J. Butte, S. J. Swamidass, and Z. Lu, "A survey of current trends in computational drug repositioning," *Briefings in bioinformatics*, vol. 17, no. 1, pp. 2–12, 2016.

[28] G. Benaim, J. M. Sanders, Y. Garcia-Marchán, C. Colina, R. Lira, A. R. Caldera, G. Payares, C. Sanoja, J. M. Burgos, A. Leon-Rossell, *et al.*, "Amiodarone has intrinsic anti-trypanosoma c ruzi activity and acts synergistically with posaconazole," *Journal of medicinal chemistry*, vol. 49, no. 3, pp. 892–899, 2006.

[29] K. E. Kinnamon, B. T. Poon, W. L. Hanson, and V. B. Waits, "Activity of anticancer compounds against trypanosoma cruzi-infected mice.," *The American journal of tropical medicine and hygiene*, vol. 58, no. 6, pp. 804–806, 1998.

[30] F. Derouin and M. Santillana-Hayat, "Anti-toxoplasma activities of antiretroviral drugs and interactions with pyrimethamine and sulfadiazine in vitro," *Antimicrobial agents and chemotherapy*, vol. 44, no. 9, pp. 2575–2577, 2000.

[31] F. Evens, K. NIBMEGEEES, A. PACKOHANIAN, *et al.*, "Nitrofurazone therapy of trypanosoma gambiense sleeping sickness in man.," *American Journal of Tropical Medicine and Hygiene*, vol. 6, no. 4, pp. 665–78, 1957.

[32] J. Alonso-Padilla and A. Rodríguez, "High throughput screening for anti–trypanosoma cruzi drug discovery," *PLoS Negl Trop Dis*, vol. 8, no. 12, p. e3259, 2014.

[33] J. C. Engel, K. K. Ang, S. Chen, M. R. Arkin, J. H. McKerrow, and P. S. Doyle, "Image-based high-throughput drug screening targeting the intracellular stage of trypanosoma cruzi, the agent of chagas' disease," *Antimicrobial agents and chemotherapy*, vol. 54, no. 8, pp. 3326–3334, 2010.

[34] L. Lucantoni, F. Silvestrini, M. Signore, G. Siciliano, M. Eldering, K. J. Dechering, V. M. Avery, and P. Alano, "A simple and predictive phenotypic high content imaging assay for plasmodium falciparum mature gametocytes to identify malaria transmission blocking compounds," *Scientific reports*, vol. 5, p. 16414, 2015.

[35] M. De Rycker, I. Hallyburton, J. Thomas, L. Campbell, S. Wyllie, D. Joshi, S. Cameron, I. H. Gilbert, P. G. Wyatt, J. A. Frearson, *et al.*, "Comparison of a high-throughput high-content intracellular leishmania donovani assay with an axenic amastigote assay," *Antimicrobial agents and chemotherapy*, vol. 57, no. 7, pp. 2913–2922, 2013.

[36] B. Institute, "Luminescence cell-based/microorganism primary hts to identify inhibitors of trypanosoma cruzi replication. pubchem bioassay aid 1885.." Available at: http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=1885, October 2009. Accessed: 01.04.2017.

[37] A. N. Lima, E. A. Philot, G. H. G. Trossini, L. P. B. Scott, V. G. Maltarollo, and K. M. Honorio, "Use of machine learning approaches for novel drug discovery," *Expert opinion on drug discovery*, vol. 11, no. 3, pp. 225–239, 2016.

[38] S. Forli, "Charting a path to success in virtual screening," *Molecules*, vol. 20, no. 10, pp. 18732–18758, 2015.

[39] P. Larranaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J. A. Lozano, R. Armañanzas, G. Santafé, A. Pérez, *et al.*, "Machine learning in bioinformatics," *Briefings in bioinformatics*, vol. 7, no. 1, pp. 86–112, 2006.

[40] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, *Proceeding of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*, ch. Supervised machine learning: a review of classification techniques, pp. 3–24. 2007.

[41] A. Lavecchia, "Machine-learning approaches in drug discovery: methods and applications," *Drug discovery today*, vol. 20, no. 3, pp. 318–331, 2015.

[42] G. B. Goh, N. O. Hodas, and A. Vishnu, "Deep learning for computational chemistry," *Journal of Computational Chemistry*, 2017.

[43] L. J. Lu, Y. Xia, A. Paccanaro, H. Yu, and M. Gerstein, "Assessing the limits of genomic data integration for predicting protein networks," *Genome research*, vol. 15, no. 7, pp. 945–953, 2005.

[44] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.

[45] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE transactions on pattern analysis and machine intelligence*, vol. 20, no. 8, pp. 832–844, 1998.

[46] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[47] Y. Freund and R. E. Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," in *European conference on computational learning theory*, pp. 23–37, Springer, 1995.

[48] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[49] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, "Random forest: a classification and regression tool for compound classification and qsar modeling," *Journal of chemical information and computer sciences*, vol. 43, no. 6, pp. 1947–1958, 2003.

[50] H. Kubinyi, "Similarity and dissimilarity: a medicinal chemist's view," *Perspectives in Drug Discovery and Design*, vol. 9, no. 11, pp. 225–252, 1998.

[51] X. Chen and C. H. Reynolds, "Performance of similarity measures in 2d fragment-based similarity searching: comparison of structural descriptors and similarity coefficients," *Journal of chemical information and computer sciences*, vol. 42, no. 6, pp. 1407–1414, 2002.

[52] P. Willett, "Similarity-based virtual screening using 2d fingerprints," *Drug discovery today*, vol. 11, no. 23, pp. 1046–1053, 2006.

[53] D. Bajusz, A. Rácz, and K. Héberger, "Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations?," *Journal of cheminformatics*, vol. 7, no. 1, p. 20, 2015.

[54] G. M. Keserü and G. M. Makara, "The influence of lead discovery strategies on the properties of drug candidates," *Nature reviews. Drug discovery*, vol. 8, no. 3, p. 203, 2009.

[55] K. Faust, D. Croes, and J. van Helden, "Prediction of metabolic pathways from genome-scale metabolic networks," *Biosystems*, vol. 105, no. 2, pp. 109–121, 2011.

[56] H. Cheng, J. Lv, H. Wang, B. Wang, Z. Li, and Z. Deng, "Genetically engineered pichia pastoris yeast for conversion of glucose to xylitol by a single-fermentation process," *Applied microbiology and biotechnology*, vol. 98, no. 8, pp. 3539–3552, 2014.

[57] J. Klekota and F. P. Roth, "Chemical substructures that enrich for biological activity," *Bioinformatics*, vol. 24, no. 21, pp. 2518–2525, 2008.

[58] N. Schneider, C. Jäckels, C. Andres, and M. C. Hutter, "Gradual in silico filtering for druglike substances," *Journal of chemical information and modeling*, vol. 48, no. 3, pp. 613–628, 2008.

[59] A. Koutsoukas, R. Lowe, Y. KalantarMotamedi, H. Y. Mussa, W. Klaffke, J. B. Mitchell, R. C. Glen, and A. Bender, "In silico target predictions: defining a benchmarking data set and comparison of performance of the multiclass naïve bayes and parzen-rosenblatt window," *Journal of chemical information and modeling*, vol. 53, no. 8, pp. 1957–1966, 2013.

[60] F. Nigsch, A. Bender, J. L. Jenkins, and J. B. Mitchell, "Ligand-target prediction using winnow and naive bayesian algorithms and the implications of overall performance statistics," *Journal of chemical information and modeling*, vol. 48, no. 12, pp. 2313–2325, 2008.

[61] S. L. Kinnings, N. Liu, P. J. Tonge, R. M. Jackson, L. Xie, and P. E. Bourne, "A machine learning based method to improve docking scoring functions and its application to drug repurposing," *Journal of chemical information and modeling*, vol. 51, no. 2, p. 408, 2011.

[62] K. Heikamp and J. Bajorath, "Prediction of compounds with closely related activity profiles using weighted support vector machine linear combinations," *Journal of chemical information and modeling*, vol. 53, no. 4, pp. 791–801, 2013.

[63] J. A. Castillo-Garit, M. C. Vega, M. Rolón, Y. Marrero-Ponce, A. Gómez-Barrio, J. A. Escario, A. A. Bello, A. Montero, F. Torrens, F. Pérez-Giménez, *et al.*, "Ligand-based discovery of novel trypanosomicidal drug-like compounds: In silico identification and experimental support," *European journal of medicinal chemistry*, vol. 46, no. 8, pp. 3324–3330, 2011.

[64] J. A. Castillo-Garit, O. del Toro-Cortés, V. V. Kouznetsov, C. O. Puentes, A. R. Romero Bohórquez, M. C. Vega, M. Rolón, J. A. Escario, A. Gómez-Barrio, Y. Marrero-Ponce, *et al.*, "Identification in silico and in vitro of novel trypanosomicidal drug-like compounds," *Chemical biology & drug design*, vol. 80, no. 1, pp. 38–45, 2012.

[65] J. A. Castillo-Garit, O. del Toro-Cortés, M. C. Vega, M. Rolón, A. R. de Arias, G. M. Casanola-Martin, J. A. Escario, A. Gómez-Barrio, Y. Marrero-Ponce, F. Torrens, *et al.*, "Bond-based bilinear indices for computational discovery of novel trypanosomicidal drug-like compounds through virtual screening," *European journal of medicinal chemistry*, vol. 96, pp. 238–244, 2015.

[66] S. Ekins, J. Lage de Siqueira-Neto, L.-I. McCall, M. Sarker, M. Yadav, E. L. Ponder, E. A. Kallel, D. Kellar, S. Chen, M. Arkin, B. A. Bunin, J. H. McKerrow, and C. Talcott, "Machine learning models and pathway genome data base for trypanosoma cruzi drug discovery," *PLOS Neglected Tropical Diseases*, vol. 9, pp. 1–18, 06 2015.

[67] N. C. for Biotechnology Information, "Luminescence cell-based/microorganism dose confirmation hts to identify inhibitors of t.cruzi replication. pubchem bioassay

database; aid=2044." Available at: https://pubchem.ncbi.nlm.nih.gov/bioassay/2044, October 2009. Accessed: 15.04.2017.

[68] V. Law, C. Knox, Y. Djoumbou, *et al.*, "Drugbank 4.0: shedding new light on drug metabolism," *Nucleic acids research*, vol. 42, no. D1, pp. D1091–D1097, 2014.

[69] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *Journal of chemical information and modeling*, vol. 50, no. 5, pp. 742–754, 2010.

[70] S. D. D. Centre, "Not all logp's are calculated equal: Clogp and other short stories." Available at: https://sussexdrugdiscovery.wordpress.com/2015/02/03/not-all-logps-are-calculated-equal-clogp-and-other-short-stories/, February 2015. Accessed: 13.04.2017.

[71] A. K. Ghose and G. M. Crippen, "Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure-activity relationships. 2. modeling dispersive and hydrophobic interactions," *Journal of chemical information and computer sciences*, vol. 27, no. 1, pp. 21–35, 1987.

[72] G. R. Bickerton, G. V. Paolini, J. Besnard, S. Muresan, and A. L. Hopkins, "Quantifying the chemical beauty of drugs," *Nature chemistry*, vol. 4, no. 2, pp. 90–98, 2012.

[73] D. F. Veber, S. R. Johnson, H.-Y. Cheng, B. R. Smith, K. W. Ward, and K. D. Kopple, "Molecular properties that influence the oral bioavailability of drug candidates," *Journal of medicinal chemistry*, vol. 45, no. 12, pp. 2615–2623, 2002.

[74] Y. Wang, S. H. Bryant, T. Cheng, J. Wang, A. Gindulyte, B. A. Shoemaker, P. A. Thiessen, S. He, and J. Zhang, "Pubchem bioassay: 2017 update," *Nucleic Acids Research*, vol. 45, no. D1, pp. D955–D963, 2017.

[75] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[76] A. R. Leach and V. J. Gillet, *An introduction to chemoinformatics*. Springer Science & Business Media, 2007.

[77] C. M. O'Connor, J. U. Adams, and J. Fairman, "Essentials of cell biology," *Cambridge: NPG Education*, 2010.

[78] U. Consortium *et al.*, "Uniprot: the universal protein knowledgebase," *Nucleic acids research*, vol. 45, no. D1, pp. D158–D169, 2017.

[79] H. Yu, N. M. Luscombe, H. X. Lu, X. Zhu, Y. Xia, J.-D. J. Han, N. Bertin, S. Chung, M. Vidal, and M. Gerstein, "Annotation transfer between genomes: protein–protein interologs and protein–dna regulogs," *Genome research*, vol. 14, no. 6, pp. 1107–1118, 2004.

[80] F. Altschul, Stephen, W. Gish, *et al.*, "Basic local alignment search tool," *J. Mol. Biol.*, 1990.

[81] N. Triballeau, F. Acher, I. Brabet, J.-P. Pin, and H.-O. Bertrand, "Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. application to high-throughput docking on metabotropic glutamate receptor subtype 4," *Journal of medicinal chemistry*, vol. 48, no. 7, pp. 2534–2547, 2005.

[82] C. Empereur-mot, H. Guillemain, A. Latouche, J.-F. Zagury, V. Viallon, and M. Montes, "Predictiveness curves in virtual screening," *Journal of cheminformatics*, vol. 7, no. 1, p. 52, 2015.

[83] S. V. Stehman, "Selecting and interpreting measures of thematic classification accuracy," *Remote sensing of Environment*, vol. 62, no. 1, pp. 77–89, 1997.

[84] M. Ringnér, "What is principal component analysis?," *Nature biotechnology*, vol. 26, no. 3, pp. 303–304, 2008.

[85] J. R. Quinlan, *C4. 5: programs for machine learning*. Elsevier, 2014.

[86] W. Iba and P. Langley, "Induction of one-level decision trees," in *Proceedings of the ninth international conference on machine learning*, pp. 233–240, 1992.

[87] B. Chen, R. P. Sheridan, V. Hornak, and J. H. Voigt, "Comparison of random forest and pipeline pilot naive bayes in prospective qsar predictions," *Journal of chemical information and modeling*, vol. 52, no. 3, pp. 792–803, 2012.

[88] W. Apt, X. Aguilera, A. Arribada, C. Pérez, C. Miranda, G. Sánchez, I. Zulantay, P. Cortes, J. Rodriguez, and D. Juri, "Treatment of chronic chagas' disease with itraconazole and allopurinol.," *The American journal of tropical medicine and hygiene*, vol. 59, no. 1, pp. 133–138, 1998.