Chemical similarity of cancer drugs and the construction of a comprehensive drug target network, focusing on the cancer metabolic pathway.

Diploma Thesis

G. Kallergis

October 23, 2019

Advisors: Prof. Zervakis Michail, Prof. Stavrakakis Georgios, Dr. Sfakianakis Stelios

Department of Electrical and Computer Engineering, Technical University Of Crete

acknowledgements

At first, I would like to thank my supervisor, Prof. Zervakis for his useful guidance and support. For their support and understanding, I would also like to thank and the rest members of my thesis committee, Prof. Stavrakakis and Dr. Stelios Sfakianakis.

Special thanks to my family, my parents Polydoros and Eleni and my sister Kallia who always help me in every way they can, standing on my side. In addition, I want to thank my friends and the people with whom I shared my office in FORTH who motivated and inspired me.

Last but not least, I would like to thank Dr. Marios Spanakis and Dr. Stelios Sfakianakis from Computational Biomedicine Lab of FORTH for their guidance and collaboration all these years, firstly in my internship and later, in my Diploma Thesis. Their contribution was crucial for this work and I was taught a lot by working by their side. They are both great scientists and kind people and I am sincerely grateful to them.

abstract

Drug Repositioning using *in silico* approaches, gains more and more ground both in academia and in the R& D departments of big pharmaceutical companies mostly because they result in a significant reduction to cost and time assisting in designing more efficient and safer drug molecules. One of the main fields it has been deployed is treating and preventing cancer, which despite the massive efforts on time and investments has not been efficiently tackled yet. This has led to reposition compounds like Acetylsalicylic acid (Aspirin) and Metformin to cure various types of tumour. Chemical similarity is an important concept in drug research. *In silico* tools for similarity-based approaches are used to identify compounds with similar bioactivities based on structural similarity between two ligands that could share same or similar targets. As chemical similarity seems to be an indication for similar pharmacological activity, this thesis aims to take advantage of it and develop two computational approaches to find approved or experimental drugs which could possibly be used to treat cancer. These are based on algorithms common in chemoinformatics, on machine learning models, and ensemble methods to enhance their performance in order to make confident predictions. The approaches are based on the assessment of the Simplified Molecular-Input Line-Entry System (SMILES) as adequate molecular structure representations for the identification of drug similarities. The first method aims to pairwise drug similarity or similarity among a few compounds, whereas the second one focuses on drug group similarity. Results include suggestions of coxib similar drugs and repositioned drugs focusing on metabolic pathways of cancer.

Contents

Chapter 1

Introduction

Discovering alternative uses for approved or experimental drugs is one of the most interesting and exciting topics in computational pharmaceutics. Drug repositioning [1], as it is called, has drawned a lot of attention due to the advantages it provides. At first, there is much lower risk of failure due to safety concerns as it has already passed some preclincal trials and in addition, demands much less time and money for drug development. Another noticeable advantage is that repositioned drugs can be used for rare diseases as the economics for developing such medication is unpropitious for a pharmaceutical company, resulting to non existent medication for many diseases of this kind. The benefits and the potential of this field steered academia and pharmaceutical companies towards this direction. As a result, novel experimental and computational approaches have been developed with the majority of them belonging to the latter one. [40]

A factor that played crucial role was the rapid increase of computational power. To be more specific, the constant enhancement of hardware, predicted in Moore's law, unleashed our capability to generate, store and process large amounts of data. A lot of research has been carried out in order to improve those procedures leading to our current position, handling a great number of data using sophisticated and computationally more demanding algorithms.

Due to this progress, there are several computational approaches in drug repositioning field such as molecular docking, genetic association and text mining in Electronic Health Records [40]. Particularly interesting is signature matching, which stands on the comparison of the unique characteristics of a drug against all the others [23]. Part of signature matching is also chemical similarity of drugs which is valuable for the area of knowledge discussed as chemically similar compounds may have the same biological or pharmacological effect [36]. This could be very beneficial for finding new medications to treat cancer. Despite the massive efforts and investment to cure this disease, the desirable results have not been yielded yet. There is still need for more effective drugs, with less side effects and drug-drug interactions. [41]

Trying to contribute on this problem, this thesis takes advantage of chemical similarity and concentrates on finding alternative uses for existing drugs with common chemical structures using algorithms and machine learning models. Hence, the repositioned compounds will be proposed as anticancer drugs, however, the models used can be applied in any category of drugs.

1.1 Chemical Representations

Finding such drugs is not an easy process and from the above statement becomes clear that the chemical representation of molecules will play crucial role when implementing the chemical representation. There are multiple ways to do so which can be separated into three main categories: 1- dimensional (1D), 2D and 3D. The last ones are far more complex, demanding much computational power and techniques such as image processing etc. The most simple format is 1D, a string containing the chemical information such as InChi, SMARTS and SMILES. For this thesis, SMILES are chosen for being a very simple but also very effective encoding [37]. It is very challenging to take the least possible information and perform well utilizing of such a representation.

1.1.1 SMILES

SMILES [53] stands for Simplified Molecular-Input Line-Entry System and it is a line notation used to encode molecular structures as strings. Those strings are generated using depth-first tree traversal of a chemical graph. There is a predefined set of rules to follow in order to get a drug's SMILES, enabling to integrate as much information as possible in the ASCII characters contained. In the following figure is demonstrated the 2D structure of Vanillin.

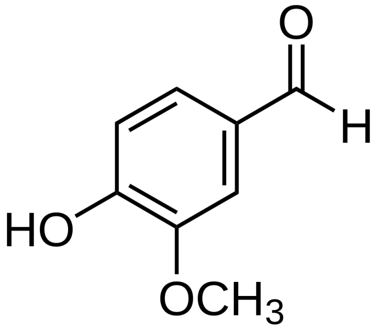


Figure 1.1: The structure of Vanilin

Its SMILES is O=Cc1ccc(O)c(OC)c1. Analyzing this string, letters in parenthesis represent branches, capital characters represent the atoms whereas, small 'c' represents the carbon atoms of an aromatic ring. The number following it, stands for the number of aromatic rings and the symbol of equal indicates the double bond. So, in our example, it becomes easy to understand that Vanillin contains an aromatic ring, 2 branches and an atom of Oxygen with a double bond. Of course, there are more representations for more complex molecules but it is not necessary for the reader to delve deeper into this.

1.2 Prior Art

A lot of research has been carried out in the field of drug repositioning. There are two main approaches here, computational and experimental and of course the one that we are interested, is the first one. Computational efforts are data-driven, deploying a wide range of tools from different fields of computer science. Drug based similarity relies on drug-drug similarity which can be expressed by similar chemical structures and also by their associations with similar targets. However, there are models combining much more types of data such as the ones mentioned before and in addition, diseases and genes [56].

Regarding chemical similarity for drug repositioning, it has been explored a lot as an option, stating the chemical structure representation, 2-d or 3-d, their prons and cons, the similarity measures, the data used and the biological problem it has been applied to [54] [47] [51] [6]. What is more, a paper of 2017, compared the chemical similarity methods for natural products[48]. There are numerous efforts utilizing molecular fingerprints which can be separated into two categories; the first one uses some kind of fingerprints and a similarity measure, usually Dice or Tanimoto and the other one uses fingerprints as feature vectors on a machine learning model. The first approach has been adopted by Xue [57]. Ozturk has also evaluated different methods for drug drug similarity like Longest Common Subsequence and different versions of it. [37]

Machine learning models have been adopted the last years with the development of those methods. Heikamp et al [19] proposed a model using MACCS and ECFP4 fingerprints in combination with KNN for k=1,5,10 neighbors. The similarity measure was Tanimoto coefficient. Riniker et al. [44] suggested similarity maps using Morgan fingerprints which can be extended to machine learning models- Random Forests and Bayes- to visualize the predicted probability of the model. Weighted similarity-based clustering using ECFP6 was applied on EGFR and FGFR projects containing genes and primary assays to determine bioactivity [39] . Another work based on this chemical representation was SubMat [48]. Two databases have been investigated (IR,MS) by deploying clustering relying on Tanimoto similarity in fingerprints. Moving forward from machine learning to deep learning which develops rapidly, Duvenaud et al. [49] proposed the neural fingerprint method as one of the first efforts in creating a graph convolution model. Bjerrum proposed to use SMILES as input for LSTM RNNs which showed that augmenting the dataset with new SMILES can lead to better results.

1.3 Thesis Outline

In **Chapter 2**, the chemoinformatics methods are utilized to extract features from SMILES that are necessary for the proposed models. In addition, the similarity metric for each method is presented.

In **Chapter 3**, we describe the machine learning models that were deployed. So, their algorithms are presented, evaluation metrics and the ensemble methods.

In **Chapter 4** describes the models that were developed and also, contains the validation of the models.

In **Chapter 5**, the experimental results for both models are provided and discussed. They are applied in real word problems using large databases and their results are analyzed and presented.

Finally, in **Chapters 6 and 7**, there are some conclusions drawn from this thesis and what is more, some suggestions for future work and possible extensions of this work.

Chapter 2

Feature Extraction from SMILES

In the following pages the algorithms used to find similar drugs are described. There are two categories: the first one is composed by algorithms treating SMILES as chemical compounds and the second one by algorithms treating SMILES as strings. Of course, in both categories there are numerous of them , however, either the best among them are chosen or the algorithms that complement each other. For instance, in the first category, each method focuses on different kind of features so all of them considered, create a more general view of the similiarity between the compounds.

2.1 Chemical features

In this section the algorithms related to the chemical representation of SMILES are presented. Even though there are several methods to approach it, fingerprints are used in this thesis. The reason behind this choice lies on the several advantages provided by this method. To be more specific, due to the complexity of this task, an abstract, simple and computationally light approach is needed. All these characterize molecular fingerprints.

Fingerprints are bitstrings which encode the presence or absence of a structure. Depending on the type of fingerprint, they might encode different patterns, atoms, bonds and more chemical features, as it will be explained in more detail in the following lines. This representation of SMILES makes it easier to compare and assess the similarity of two different compounds using the appropriate metric. Moreover, fingerprints can be used as features in machine learning models as it will also be analyzed later.

The fingerprints utilized here are topological, Morgan and RDKit. All of them were generated using the RDKit Open-Source Cheminformatics Software.

2.1.1 Topological Fingerprints

These fingerprints are among the oldest ones, proposed in 1986 [35]. This fingerprint uses a representation of the sequence of four atoms, taking into account their atomic type, the number of π electrons pairs and the non-hydrogen branches, assigning it to the path-based fingerprints. Trying to make it easier, the authors use the following format for each non hydrogen atom: (NPI-TYPE-NBR), where NPI stands for the number of π electrons on each atom, TYPE relates to the atomic species and NBR, to the Non Hydrogen Branches. So, having four them it ends up like this: (NPI-TYPE-NBR)-(NPI-TYPE-NBR)-(NPI-TYPE-NBR)-(NPI-TYPE-NBR)

An example of the topological torsion is shown below:

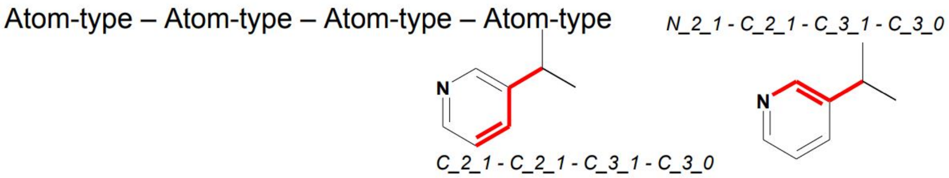

Atom-type – Atom-type – Atom-type – Atom-type
N_2_1 - C_2_1 - C_3_1 - C_3_0
N
N
C_2_1 - C_2_1 - C_3_1 - C_3_0

Figure 2.1: Topological Fingerprint according to (NPI-TYPE-NBR). Figure used by G Laundrum's presentation,"Fingerprints in the RDKit",Novartis Institutes for BioMedical Research, Basel, RDKit UGM 2012, London

2.1.2 Morgan Fingerprints

Morgan or Circular Fingerprints shows another version of ECFP fingerprints [33]. They are similarity fingerprints taking into account the neighborhood of the atom which depends on the radius provided by the user. The process for their generation is as follows:

1. Assign each atom with an identifier.
2. Update iteratively each atom's identifiers taking into account its neighbours and hash.
3. Remove duplicates.
4. Fold list of identifiers into a 2048-bit vector (a Morgan fingerprint).

Figure 2.2 shows how the radius affects the Morgan fingerprint. Comparing those two, it is apparent that the greater the radius is, the more atom and bond types are included.

2.1.3 RDKit Fingerprints

This is a fingeprint based on Daylight Fingerprint, provided by RDKit Chemoinformatics toolkit. It is a substructure fingerprint, generating hashed molec-

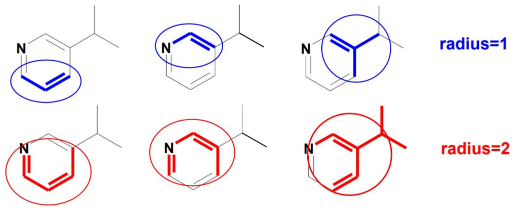


Figure 2.2: Representation of the atoms and bonds included in Morgan fingerprint for radius equal to 1 and 2. Figure used by G Laundrum's presentation," Fingerprints in the RDKit",Novartis Institutes for BioMedical Research, Basel, RDKit UGM 2012, London

ular subgraphs. Hash for the path is based on bond types and each bond's neighbor count. The bond types mentioned earlier are set by atom and bond types and atom types are set by atomic number and aromaticity.

The algorithm functions by finding all subgraphs between minPath and maxPath in length. For each subgraph:

1. Generate hash for the path using bond types and each bond's neighbor count
2. Seed random-number generator with hash
3. Generate as many random numbers as the number of bits per hash between 0 and fingerprint's size and set the corresponding bits in the fingerprint.

The algorithm is described according to RDKit Open-Source Cheminformatics Software's documentation.

2.1.4 Similarity Measurement of Fingerprints

There is a great variety of metrics to measure the similarity of two compounds and there has been conducted research comparing them. Among the most prominent are Tanimoto, Tversky and lots more more but the one used here is the Dice Index. This is used to compare Fingerprints and TF-IDF on Lingos.

Dice Similarity

Dice similarity for dichotomous variables is expressed as below:

$$Dice\_index = \frac{2 * c}{a + b}$$

Where c corresponds to the common features of the two drugs compared, a equals to the features present only in drug-A and b to drugs compared only to features of drug- B. The threshold to determine if they are similar was set to be 0.5. [11][49]

2.2 String Features

This kind of algorithms almost ignores the chemical significance of SMILES and instead, they treat it as a simple string. Considering them so, is a key move as it enables to use a lot of algorithms and techniques from Information Retrieval field and explore in a different way. Therefore, many methods have been used such as Edit and City Block Distance [37] . Some of those are Term Frequency- Inverse Document Frequency (TF-IDF) on LINGOS and Maximal Common Subsequence deployed here.

2.2.1 Maximum Common Subsequence (MCS)

MCS [22] was used in order to extract knowledge from the SMILES as a string. This method finds the common subsequence which is compared to MCS which just takes the longest one. Its application lies on the fact that compounds with similar subsequences could possibly have the same chemical properties as mentioned in introduction. Therefore, they should have a common consecutive subtructure in their SMILES and not just common fragments all over the string. To achieve this, the MCS of drugs is found and then it is used as a lead to find the MCS with the drug candidate. Based on the MCS the similarity is calculated as:

$$MCS(S_i, S_j) = \frac{length(MCS(S_i, S_j))^2}{length(S_i) \cdot length(S_j)},$$

Chapter 3

Machine Learning Models

3.1 Data Processings

The data used usually in such applications are scarce and the classes are highly imbalanced resulting to difficulties and suboptimal use of machine learning models. Thus, the data have to be processed a bit at first, by applying specific techniques as the one following and careful collection of data.

After that processing, a fingerprint is used as input for the machine learning models. The bits of a fingerprint can be used as feature vector for the machine learning model. Each feature represents the presence or the absence of a chemical structure in the drug. The models use them to get trained and then predict similar compounds.

3.1.1 SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) is a method creating samples for the minority class through oversampling.[7] In order to achieve this, it utilizes all the samples of this class and adds new, synthetic ones on the lines joining this sample with some of its k nearest neighbors which also belong in the same class. SMOTE algorithm chooses randomly the point where the new sample will be generated on the line mentioned before. In addition, the neighbors that will be used are chosen randomly; depending on the number of samples needed, some of the k closest neighbors are selected. Therefore, the synthetic samples cause the classifier to create larger and less specific decision regions, enabling it to generalize better.

Figure 3.1 shows how SMOTE algorithm works. Obviously, there is a minority class represented by green dots whereas the majority class is represented by grey squares. SMOTE finds the 3 nearest neighbors (x2,x3,x4) of sample x1 and generates the synthetic samples on random positions on the lines joining the sample with its neighbors (red diamonds).

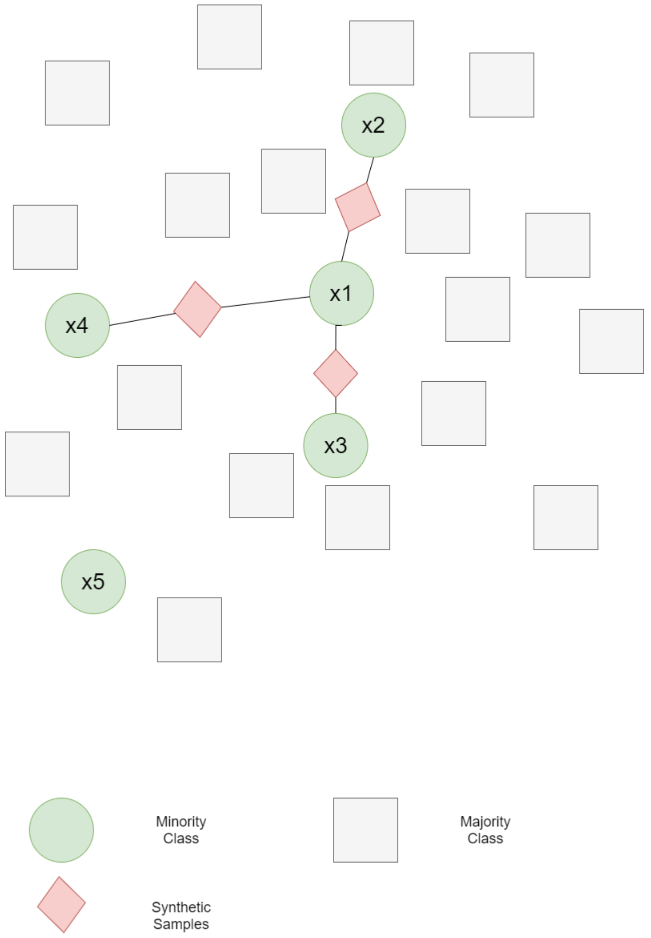


Figure 3.1: Representation of SMOTE algorithm

3.2 Supervised Learning Algorithms

3.2.1 Logistic Regression

It is a statistical model using logistic function to model a dataset containing one or more independent variables.[13] As a supervised model used for binary classification , the output is either 1 or 0. Logistic Regression is a method for fitting a regression curve, y=f(x), where f is the logistic function. A simple sigmoid function is the logistic function:

$$y = \frac{e^x}{1 + e^x}$$

with y being the outcome (0,1) and x the input variables.This can be extended to

$$y = \frac{e^{ax+b}}{1 + e^{ax+b}},$$

where a and b are the logistic coefficients and stand for logistic intercept and slope.

As a result, given the input x, sigmoid tries to adjust the sigmoid function and classify the data.

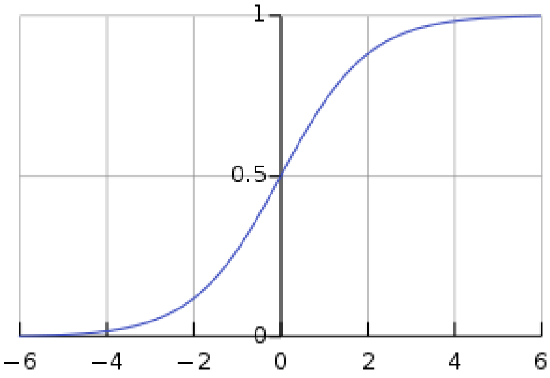


Figure 3.2: The logistic function

3.2.2 k- Nearest Neighbors (k-NN)

KNN is a non parametric, instance based learning model [10]. It is distribution free, stores the training set as a representation, so there is not actual training in the way it is defined in other models and is lazy, meaning it does not generalize. A characteristic of this model is the calculation of distance, usually Euclidean, to determine the feature distance and get the nearest k samples. Majority Voting is used to determine the final label of the recently added sample.

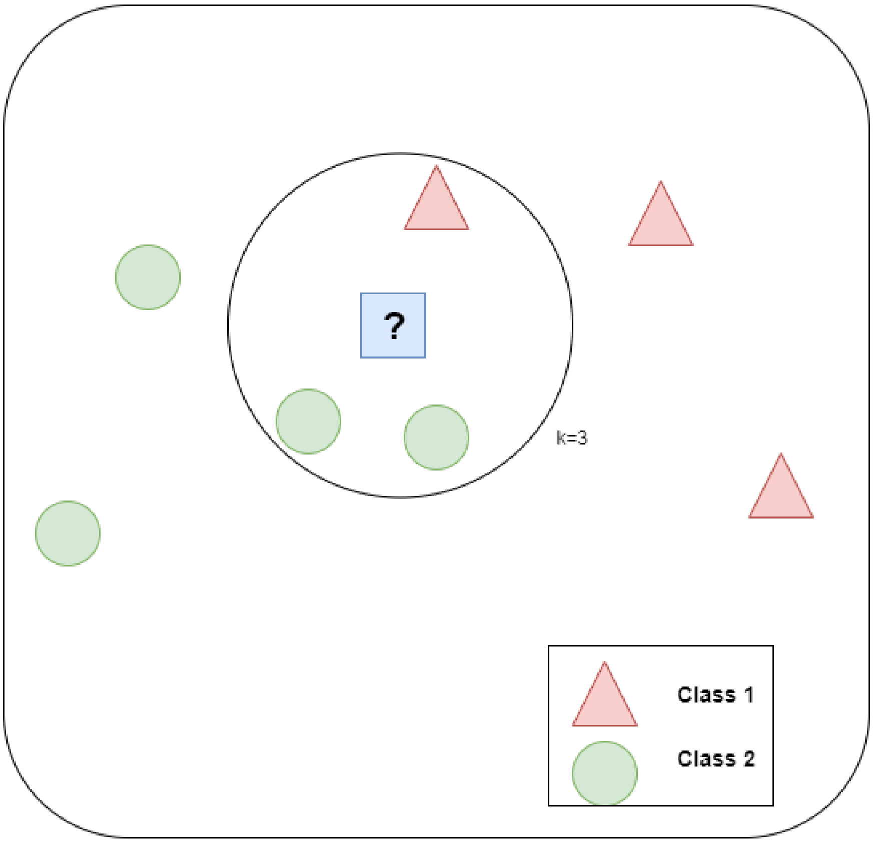


Figure 3.3: Classifying the green dot with KNN

3.2.3 Support Vector Machines (SVM)

SVM is a supervised learning classifier which given labeled training data outputs an optimal hyperplane which can classify new samples.It is utilized as a non linear mapping of input vectors into a high dimensional feature space.[9] This hyperplane is an N-dimensional space where N equals to the number of features. Utilizing support vectors, the critical elements which would affect the position of hyperplanes, tries to maximize the margin among them. One of the advantages of this model is that it can be used for not linearly separable data by using polynomial or radial basis function. Radial basis function can be described as $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|), \gamma > 0$ [38]

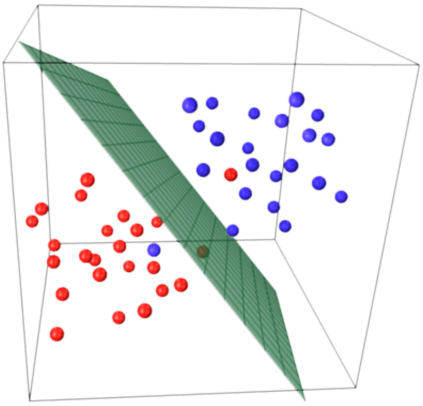


Figure 3.4: Representing the hyperplane of SVM, classifying the data. Figure obtained by [58]

3.2.4 Decision Trees

Decision Trees are a supervised algorithm which uses a tree-like graph to model of decisions. [26] Each internal node represents a "test" on an attribute such as if the coin flip was heads, each branch represents the result of this test (yes or no for the previous example), and each leaf node represents the final class label. The paths from root to leaf represent classification rules. It works for both categorical and continuous input. Below we show an example of how a decision tree looks like.

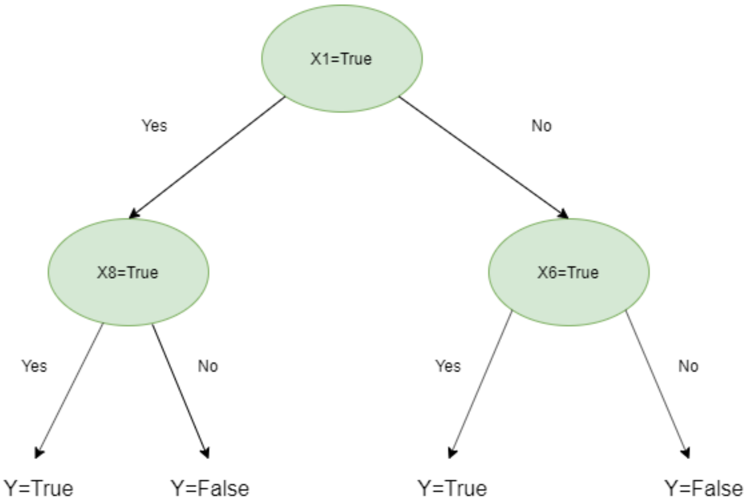


Figure 3.5: Representation of a Decision Tree

3.2.5 Random Forests

Random Forests is composed of decision trees mentioned before [4]. Each one of them makes a prediction and then, the final outcome of the model is the prediction with the most votes. The training data of each tree is sampled from the original dataset with sampling with replacement. Each tree is grown without pruning. They rely on Bagging, with main difference being the randomized feature selection which takes place in split selection by selecting a subset of features. Essentially, it enables weak or weakly-correlated classifiers to form a strong classifier.

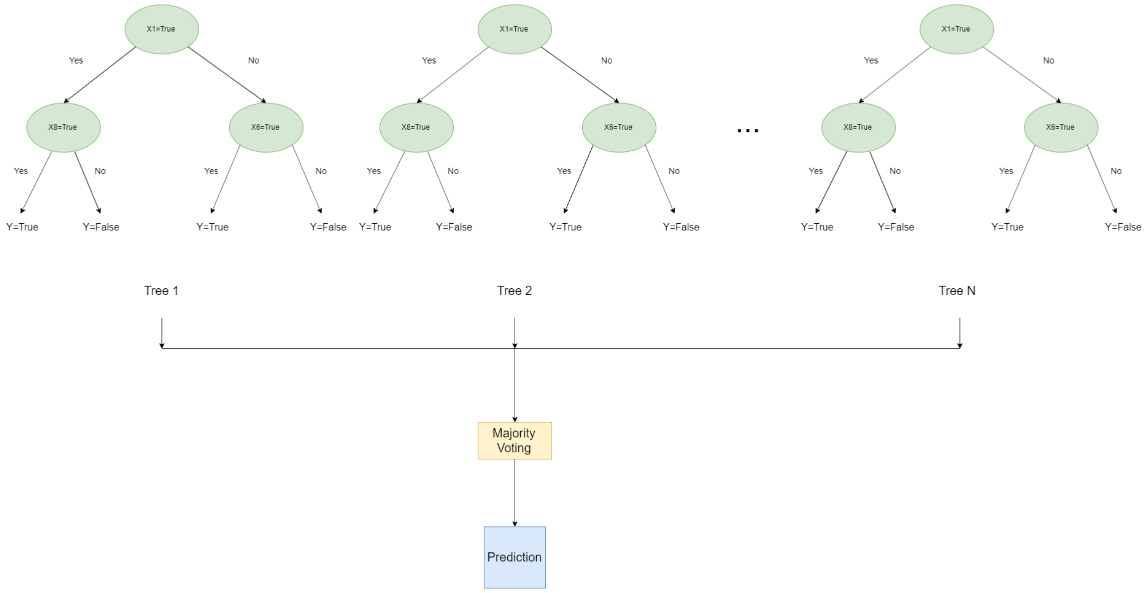


Figure 3.6: Representation of Random Forests

3.2.6 Bayes

Bayes classifier is a conditional probability model.[46] For a feature vector $x = (x_1, ..., x_N)$, where N equal to the number of features assigning it to instance probabilities and C_k equals to k-th class.

$$p(C_k|x_1, ..., x_N)$$

which can be rewritten using Bayes theorem as:

$$p(C_k|x_1, ..., x_N) = \frac{p(x_1, ..., x_N|C_k) * p(C_k)}{p(x_1, ..., x_N)}$$

Having this as basis, the classification is made using Maximum A Posteriori decision (MAP).

$$\hat{y} = argmax_{1...K} p(C_k) \prod_{i=1}^{n} p(x_i|C_k)$$

3.3 Combining different models and algorithms

3.3.1 Averaging

Averaging is one of the most simple ways to combine methods by taking as input their numerical results and outputting a real valued variable [59]. There are multiple ways to average and in the following lines two of them will be presented: simple and weighted.
For the formulas defined below, 'V' is the Value of the averaging method, $'v_i'$ the method, 'k' the number of these methods, $'w_i'$the weights of corresponding to the i-th method and 'x', the instance for which the values are calculated

Simple Averaging

This method just averages the outputs of the algorithms deployed. It is very popular and simple and one of its advantages is the real valued output which can be used along with a threshold and provide better results.

$$V(x) = \frac{1}{k} * \sum_{1}^{k} (v_i(x))$$

Weighted Averaging

Another useful method in ensemble is weighted average. Instead of having a number of classifiers or algorithms contributing the same to the final result, each prediction has a different weight. Specifically, that means that a more 'suitable' classifier should have been assigned a bigger weight. For instance, as it will be demonstrated later in more detail, when looking for compounds with aromatic rings, Morgan fingerprint should play more crucial role. As for the calculation of these weights, there are many ways to achieve it. Stochastic gradient descent, genetic algorithms, even regarding the contribution of each has been proposed. In this work, genetic algorithms have been preferred.

$$V(x) = \sum_{1}^{k} (w_i * v_i(x))$$

In addition, w_i should satisfy the following constraints:

$$w_i \geq 0 \quad and \quad \sum_{1}^{k} w_i = 1$$

3.3.2 Voting

Voting is about using classifiers or algorithms to predict the class label and that predictions is considered as a vote for the class label.[59] The final result is the class with the most votes. There are numerous techniques to reckon with those votes.

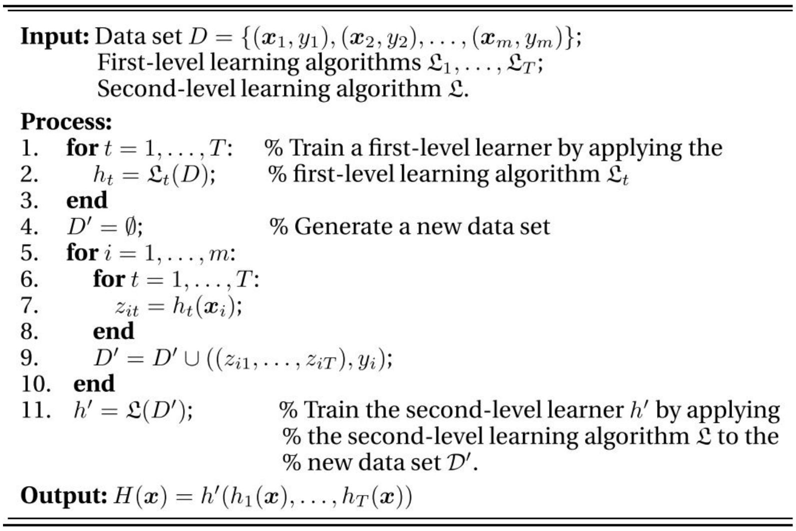
Majority Voting

Majority Voting is the most well known voting model. In this, each model outputs a class value and the class receiving more than half of the votes, is the winner. If this does not apply, then the classifier makes no prediction.

3.3.3 Ensemble using meta learners

Stacking

Stacking is a model which uses a meta-learner to combine more, individual learners [55, 3]. The individual learners are also called first -level learners and the meta-learner, second-level learner. The output of the first level is considered as input features for the next level. For its training, is essential to utilize part of the original dataset for training the first level learners and then, use the rest of the dataset for training the second level, aiming to avoid overfitting. Stacking can be either heterogenous, using different learners on its base or homogenous, having the same classifiers . [59]



```
Input: Data set $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$;
       First-level learning algorithms $\mathfrak{L}_1, \ldots, \mathfrak{L}_T$;
       Second-level learning algorithm $\mathfrak{L}$.
Process:
for $t = 1, \ldots, T$:        % Train a first-level learner by applying the
    $h_t = \mathfrak{L}_t(D)$;         % first-level learning algorithm $\mathfrak{L}_t$
end
$D' = \emptyset$;                % Generate a new data set
for $i = 1, \ldots, m$:
    for $t = 1, \ldots, T$:
        $z_{it} = h_t(\boldsymbol{x}_i)$;
    end
    $D' = D' \cup ((z_{i1}, \ldots, z_{iT}), y_i)$;
end
$h' = \mathfrak{L}(D')$;          % Train the second-level learner $h'$ by applying
                         % the second-level learning algorithm $\mathfrak{L}$ to the
                         % new data set $\mathcal{D}'$.
Output: $H(\boldsymbol{x}) = h'(h_1(\boldsymbol{x}), \ldots, h_T(\boldsymbol{x}))$
```

Figure 3.7: Algorithm of stacking [59]

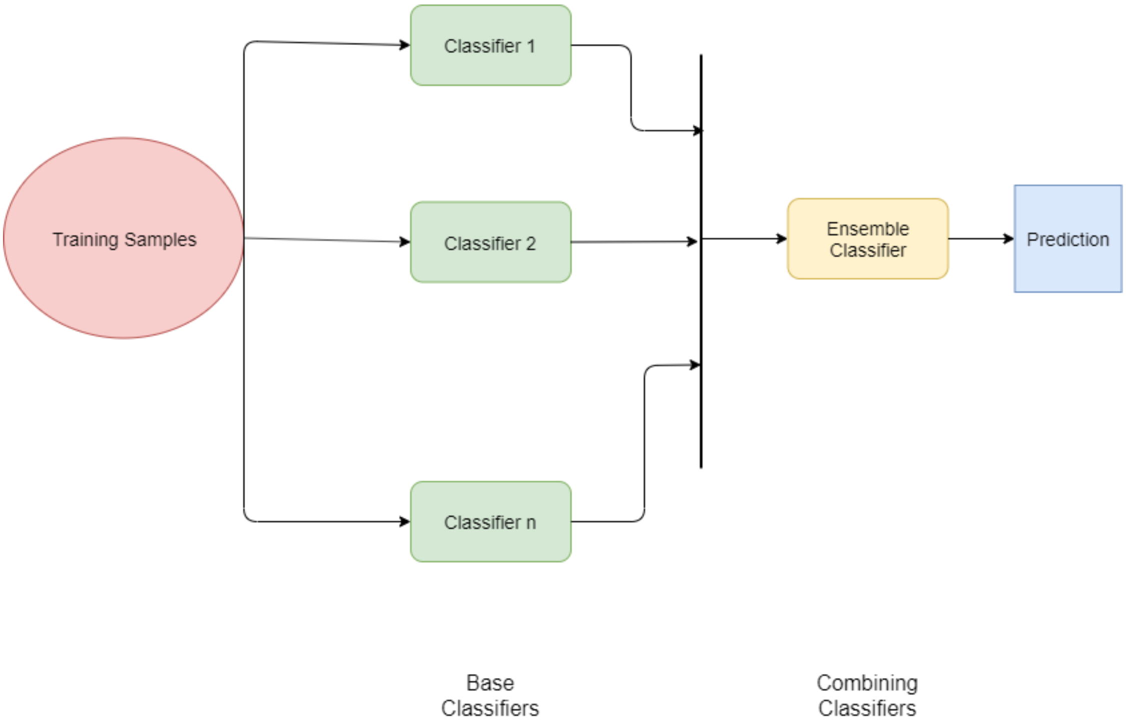


Figure 3.8: Representation of stacking

Bagging

One of the meta-algorithms used in ensemble is Bagging which is short for Bootstrap Aggregating. Its main advantage is that produces more accurate results, with reduced variance and hence, reduces the the danger of overfitting. [2] The procedure used by the algorithm is described below and there are two parts, bootstrap and aggregating. Bootstrap generates m training sets of size k from an initial dataset D with size n, where $k < n$. Those sets are formed by selecting samples uniformly with replacement. Then, m classifiers are deployed, each one trained with one of the new datasets created. Then, majority voting is applied over the results of classifiers, canceling the effect of variation[27].

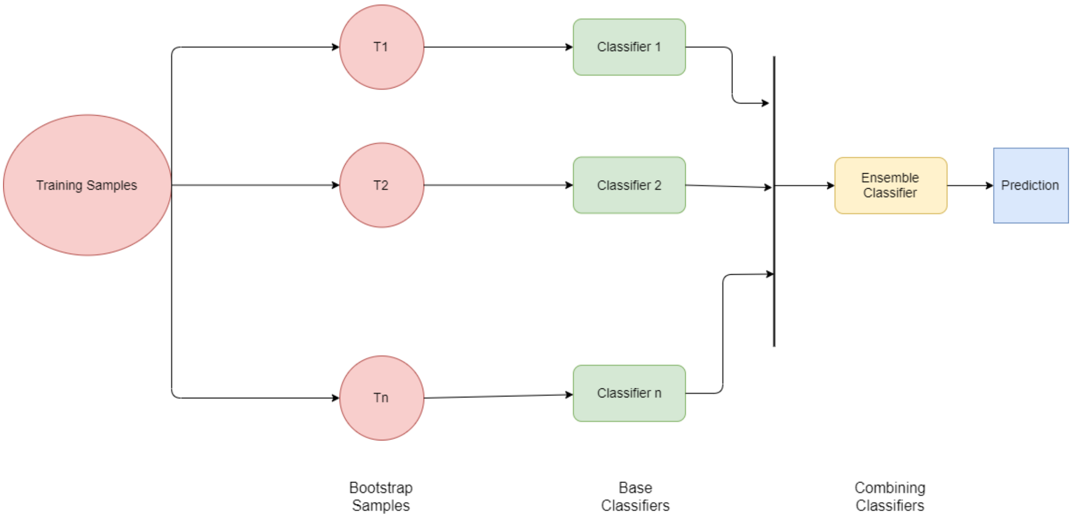


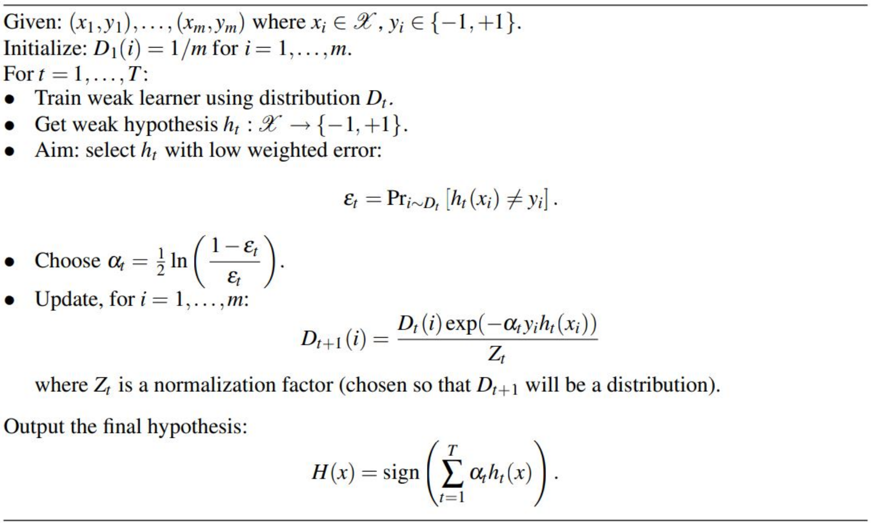
Figure 3.9: Representation of bagging

Boosting

Boosting is the last meta-algortihm of this category. Its primary goal is to reduce bias and variance by utilizing weak learners. Those are supposed to compose a stronger one through their combination. Key characteristic of this method is the sequential learning of the models instead of the parallel. The benefit is that the next model is trained on the regions that the previous learner performed poorly. So, weak learners are added to a final strong classifier by trying to reduce their errors. [59]

ADABoost

This is the most used boosting algorithm proposed by Schapire and Freund [16]. The difference of Adaptive Boosting compared to the general term of boosting algorithms relies on the term adaptive which practically means reweighing the contribution of each sample focusing on the misclassified ones. In addition, it uses all the training data rather than random subsamples of the dataset [14]. The pseudo code of the algorithm is demonstrated below as it was presented by Shappire [45].



```
Given: $(x_1, y_1), \ldots, (x_m, y_m)$ where $x_i \in \mathscr{X}$, $y_i \in \{-1, +1\}$.
Initialize: $D_1(i) = 1/m$ for $i = 1, \ldots, m$.
For $t = 1, \ldots, T$:
• Train weak learner using distribution $D_t$.
• Get weak hypothesis $h_t : \mathscr{X} \to \{-1, +1\}$.
• Aim: select $h_t$ with low weighted error:
      $\varepsilon_t = \Pr_{i \sim D_t}\left[h_t(x_i) \neq y_i\right]$.
• Choose $\alpha_t = \frac{1}{2} \ln\left(\frac{1 - \varepsilon_t}{\varepsilon_t}\right)$.
• Update, for $i = 1, \ldots, m$:
      $D_{t+1}(i) = \frac{D_t(i) \exp(-\alpha_t y_i h_t(x_i))}{Z_t}$
  where $Z_t$ is a normalization factor (chosen so that $D_{t+1}$ will be a distribution).
Output the final hypothesis:
      $H(x) = \operatorname{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$.
```

Figure 3.10: Pseudo code describing AdaBoost

3.4 Evaluation

Assessing the efficiency of the models, was of paramount importance. Having so few samples, even if later were enriched with new ones, did not steer clear of the danger of overfitting. Thus, a lot of metrics were used and techniques in order to be sure that there will not be such a problem.

3.4.1 Performance Measures

Choice of metrics clearly affects how the performance of machine learning models is evaluated and compared. Multiple of them have been suggested trying to find the best one for each case. In the following lines will be presented some of them but at first, there should be given some terms which will be useful to define the metrics.[20] [24]

True Positive (TP): Positives samples classified as positive.
True Negative (TN): Negative samples classified as negative.
False Positive (FP): Negative samples classified falsely as positive.
False Negative (FN): Positive samples classified falsely as negative.

True Positive Rate (TPR) or Sensitivity: measures the proportion of actual positives that are classified as such.
True Negative Rate (TNR) or Specificity: measures the proportion of actual negatives that are classified as such.
False Positive Rate (FPR): measures the proportion of negatives that are wrongly classified as positives.
False Negative Rate (FNR): measures the proportion of positives that are

wrongly classified as negatives.

Precision is the same as TPR and equals to the number of samples correctly classified as positive over the total number of samples classified as positive.

$$Precision = \frac{TP}{TP + FP}$$

Recall equals to the number of samples correctly classified as positive over the total number of actual positive samples.

$$Recall = \frac{TP}{TP + FN}$$

Accuracy

Accuracy is the number of correctly classified samples over all the predictions made.

$$Accuracy = \frac{TN + TP}{TN + TP + FP + FN}$$

AUC

AUC stands for Area Under the (ROC) Curve which is also demonstrated in the following figure. Its values range from 0, for an insufficient model, to 1, for a good one. It is considered to be a very good metric as it is scale invariant and classification-threshold-invariant.

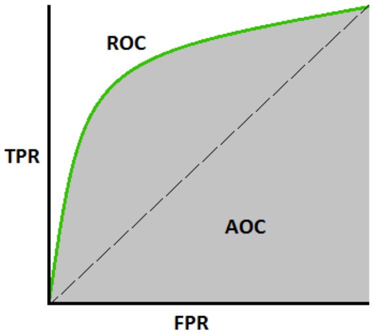


Figure 3.11: AUC on gray and ROC curve

F1- score

It is an harmonic mean of precision and recall, ranging from 0 to 1, often used when the positive class is too small.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

3.4.2 Validation

For validating the model, at first, the dataset was split in a stratified way, which means that each sample test has samples from different classes in the same proportions as in the original class.[29] Training data was composed by 70 % of the initial dataset and test data the remaining 30%. Then, stratified 10-fold validation was used in the training set, using the metrics mentioned above. The next step was to train the model and assess it again in the test set. The thought behind this was that this could help to evaluate the model, as it would provide a final objective idea of how well it performs.This is also very common when trying to check if a model overfits and generally speaking be sure that it performs well and properly. [24]

Chapter 4

Proposed Models

In this thesis, two computational approaches were developed, one for pairwise drug similarity and one more for drug group similarity, both utilizing SMILES and exploiting information deriving from it. The first one focuses on similarity between drugs using simple metrics and the second one drug group similarity, whether this drug could have the same effect as the ones belonging in the group.

4.1 Data

The data used in this thesis come from 2 databases. The first one is DrugBank, which contains more than 11000 drugs and valuable information about them. This information includes SMILES that are fundamental for our approach, drug targets associations and interactions and the whether a drug is experimental, approved, investigational and many more. Last but not least, it is essential to mention ATC codes, as they enabled to categorize drugs and obtain data with chemical or biological similarity -the fourth level of ATC indicates chemical/therapeutic/pharmacological subgroup.

The second one is Small Molecule Pathway DataBase (SMPDB). From that database were utilized the pathways and the targets included in them. To be more specific, it enabled to find targets in metabolic pathways and combined with DrugBank's knowledge to create drug-target associations.

In order to categorize the type of the pathways when analysing results, Kyoto Encyclopedia of Genes and Genomes (KEGG) was also used.

4.2 A model for drug drug similarity

This model focuses on finding similar drugs using as input the SMILES of two drugs. Those are utilized by applying the methods of chapter 2.

Regarding the fingerprints, they are produced and then compared using Dice Similarity, with the result being in range (0,1). So, there are three different similarity values from fingerprints and what is more, the outcome of Maximum Common Consecutive Substructure. The threshold of scores to be considered as such is at least 0.5. An example of this is the comparison between Fenoprofen and Flurbiprofen.

Drug 1	Drug 2	Morgan Fing.	Topological Fing.	RDKit Fing.	MCCS
Fenoprofen	Furbiprofen	0.752	0.618	0.703	0.551

Table 4.1: Results of comparing two drugs

Each of these can be used independently and show us how similar two drugs are, however, in this thesis they are combined to exploit the advantages of each one of them.

The next step is to improve their performance by combining the different perspectives of each method and have a more general view of the similarity between the compounds. This has been achieved in three ways by implementing majority voting, equal weights and weighted averaging. In the first one, if at least 3 out of 4 votes are in favor of similarity, those drugs will be considered as similar. But what if one method considers the drugs as similar, with a great percentage of similarity but not all the others, even for values being a little below threshold. By deploying equal weights averaging, the similarity score of each method is multiplied with the same weight and if the result is greater than 0.5, then they are considered as similar. In essence, if there is a very confident similarity score, it gives model the chance to consider them as similar. Apart from that, there are some methods which are considered to be more appropriate for some types of compounds. In our case, Morgan fingerprints are more suitable for coxibs in which the model will be applied. As a result, it has a greater weight than the others in order to affect more the final results.

4.2.1 Validation

The models were validated using data from DrugBank, organized in drug groups with known chemical similarity. For this purpose Non Steroid AntiInflammatory Drugs (NSAIDS), benzodiazepines, opioids, tricyclic antidepressants and corticosteroids were used. As a result in order to validate the model, drugs from those categories where chosen and it was tested if the model could successfully recognise the compounds belonging in the same drug group as similar.

Below some compounds are indicated alongside with the accuracy they had in recognizing the similar compounds.

Drug	Drug Group	Accuracy	Recall
Oxaprozine	NSAIDS	0.863	0.5
Dexibuprophen	NSAIDS	0.36	0.667
Clonazepam	Benzodiazepines	0.636	0.889
Midazolam	Benzodiazepines	0.52	0.88
Codeine	Opioids	0.704	0.8
Levorphanole	Opioids	0.63	0.8
Butriptyline	TCA	0.59	0.66
Amitriptyline	TCA	0.863	0.667

Table 4.2: Accuracy and Recall Scores for drugs of different categories with MCS method.

Drug	Drug Group	Accuracy	Recall
Oxaprozine	NSAIDS	0.63	0.5
Dexibuprophen	NSAIDS	0.227	1.0
Clonazepam	Benzodiazepines	0.52	1.0
Midazolam	Benzodiazepines	0.40	1.0
Codeine	Opioids	0.25	1.0
Levorphanole	Opioids	0.681	1.0
Butriptyline	TCA	0.431	0.66
Amitriptyline	TCA	0.47	1.0

Table 4.3: Accuracy and Recall Scores for drugs of different categories with Topo fingerprint.

Drug	Drug Group	Accuracy	Recall
Oxaprozine	NSAIDS	0.88	0.16
Dexibuprophen	NSAIDS	0.931	0.5
Clonazepam	Benzodiazepines	0.97	0.88
Midazolam	Benzodiazepines	0.933	0.66
Codeine	Opioids	0.954	0.2
Levorphanole	Opioids	0.909	0.2
Butriptyline	TCA	0.818	0.33
Amitriptyline	TCA	0.909	0.66

Table 4.4: Accuracy and Recall Scores for drugs of different categories with RDKit fingerprint.

As it becomes apparent, the weighting average model performs better than the others and this probably lies on the fact that Morgan Fingerprints play an important role in the final decision. The following table, demonstrates the success in identifying the similar compounds.

From these results, the reader can assume that weighting averaging in many

Drug	Drug Group	Accuracy	Recall
Oxaprozine	NSAIDS	0.954	0.66
Dexibuprophen	NSAIDS	0.977	0.833
Clonazepam	Benzodiazepines	0.931	1.0
Midazolam	Benzodiazepines	0.954	1.0
Codeine	Opioids	0.977	1
Levorphanole	Opioids	0.977	1
Butriptyline	TCA	0.86	1
Amitriptyline	TCA	0.886	1

Table 4.5: Accuracy and Recall Scores for drugs of different categories with RDKit fingerprint.

Drug	Majority Voting	Weighting Averaging	Equal Weights
Oxaprozine	0.5	0.833	0.833
Dexibuprophen	0.833	0.833	0.833
Clonazepam	1.0	1.0	1.0
Midazolam	0.88	0.88	1.0
Codeine	1.0	1.0	1.0
Levorphanole	0.8	1.0	1.0
Butriptyline	0.66	0.86	1.0
Amitriptyline	0.83	0.91	1.0

Table 4.6: Recall Scores for drugs of different categories for the averaging and voting methods.

Drug	Majority Voting	Weighting Averaging	Equal Weights
Oxaprozine	0.931	0.977	0.977
Dexibuprophen	0.977	0.977	0.977
Clonazepam	1.0	1.0	1.0
Midazolam	0.88	0.88	1.0
Codeine	1.0	1.0	1.0
Levorphanole	0.8	1.0	1.0
Butriptyline	0.66	0.86	1.0
Amitriptyline	0.83	0.91	1.0

Table 4.7: Accuracy Scores for drugs of different categories for the averaging and voting methods.

cases recognizes the same percentage of similar compounds. However, it has less false positive results, recognizing less drugs from other groups as similar.

4.3 A model for drug group similarity

The intuition behind this approach is to exploit chemical similarity deriving from fingerprints in order to train a machine learning model which will be able to find drugs-candidates. To be more specific, having generated

Table 4.8: Average F1 in 10-fold before and after oversampling

Model	Before	After
KNN	0.747	0.693
RF	1	0.583
LR	1	0.633
Naive Bayes	0.905	0.769
SVM	0.985	0.683
ADA	0.883	0.596

Morgan fingerprints from SMILES, these could be the input feature vector to train the model. It is not a novel approach as it has already been utilized to predict biological or pharmacological activity, something mentioned in the introduction of this thesis. Having 2048 features from each fingerprint and enough drugs from the drug group that we are interested in, will be sufficient to have an efficient model.

At first, a number of models were deployed to check their efficacy on the class imbalanced dataset. This was created by including anticancer drugs based on their ATC code that are known for having targets in metabolic pathways, thus, being the positive samples. On the contrary, anticancer with no targets in that kind of pathways and drugs not being anticancer or metabolic related, were used as negative samples. The number of the latter mentioned is much bigger, thus creating the forementioned imbalance.

The models were Logistic Regression, Support Vector Machines, Decision Trees, Random Forests, k Nearest Neighbors and Naive Bayes Classifier. Those are the most common supervised learning algorithms used in similar problems. All of them were optimally tuned by using Grid search to find the optimal parameters. Having done that each algorithm, the next step was to validate them using multiple metrics – Accuracy, Area Under Curve (AUC) and F1. In addition, the dataset was split into training and testing data (70% training and the rest testing) from which, in the first one, Stratified 10-Fold Validation was used to gain an insight into the performance of the model and afterwards, the training set was used to train the models and check if they can predict correctly the test data.

At this point, the performance in training and testing has to be evaluated but focusing more on the performance in testing. After evaluating them, even though AUC and Accuracy scores were high enough, F1 score revealed that class imbalance had an impact. Generally speaking, F1 is more appropriate to evaluate classifiers dealing with class imbalance and this is why we rely on it in the following steps. Overall, even if the results in cross validation are promising, the results in testing shows that more can be done in order the models to be able to distinguish effectively the difference between the desired and non desired drugs. The most obvious approach to potentially

Table 4.9: Classifiers' scores in 10fold validation

Model	Avg. Accuracy	Avg. AUC	Avg. F1
KNN	0.928	0.998	0.933
RF	1	1	1
LR	1	1	1
Naive Bayes	0.975	0.985	0.974
SVM	0.996	1	0.996

Table 4.10: Classifiers' scores in test set

Model	Accuracy	AUC	F1
KNN	0.890	0.959	0.905
RF	0.976	0.995	0.973
LR	0.991	0.999	0.992
Naive Bayes	0.967	0.977	0.965
SVM	0.991	0.998	0.992

provide better generalization is fixing class imbalance since, the positive samples are very few to train a model.

These scores had very low values indicating the need to have more positive samples, or less negative ones. Since the dataset is already small, undersampling could result to loss of information and generalization explaining why oversampling was preferred. So, new positive samples were generated by SMOTE algorithm, balancing the classes. As it becomes evident in table 4.8, the F1 score alongside with AUC were enhanced to a great degree.

The reader might observe that despite having a 2048 long feature vector, it has not been applied dimensionality reduction or feature selection. Some of the fore-mentioned models like Decision Trees and Random Forest select features on their own and in SVM it is not essential. However, an implementation of PCA was carried out selecting the features explaining 90% of the variance but the results did not change enough (less than 1%) to adopt and present it in this thesis.

Even though, their performance is good enough, still can be improved by using meta-learners and ensemble methods. Boosting, Bagging, Stacking and Ensemble deployed using as base estimators, SVM, Logistic Regression, Random Forests and Naive Bayes Classifier.

Comparing the results on the same dataset, the Ensemble method was preferred for having the best AUC score and overall, great scores in almost all categories for both training and testing.

Table 4.11: Ensemble's scores in 10fold validation

Model	Avg. Accuracy	Avg. AUC	Avg. F1
ADA	0.982	0.999	0.982
Ensemble	1	1	1
Stacking Classifier KNN	0.997	0.99	0.997
Stacking Classifier LRR	0.996	1	0.996
Bagging	0.994	0.999	0.995

Table 4.12: Ensemble' scores in test set

Model	Accuracy	AUC	F1
ADA	0.979	0.996	0.979
Ensemble	0.976	1	0.973
Stacking Classifier KNN	0.991	0.994	0.992
Stacking Classifier LRR	0.991	0.995	0.992
Bagging	0.979	0.999	0.979

Chapter 5

Results

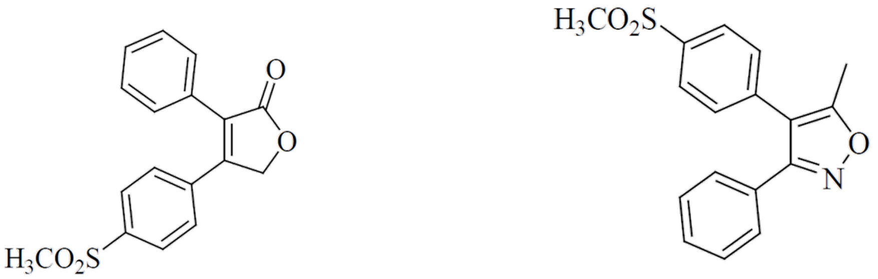
In this chapter, we present the results deriving from the application of the proposed models in two problems: finding similar drugs to coxibs, which are very interesting compounds for their anticancer actions and finding compounds used as anticancer drugs focusing on metabolic pathways. Both are very interesting problems but before moving to them, the datasets should be presented.

At first, DrugBank was utilized in finding drugs and relative information. This includes SMILES, ATC code and drug groups. Obviously, SMILES was used as input in order to generate fingerprints and be used for MCS. The ATC code allows to categorize drugs based on their recommendation for use. In addition, drug group shows if a drug is approved, investigational, experimental, nutraceutical, withdrawn or even, if it is vet approved. Apart from that, SMPDB is deployed to find targets participating in metabolic pathways. Filtering such pathways leads to get their targets which are also found in DrugBank in order to get the drug-target associations.

5.1 Model 1: An application in Coxibs

As it was stated in introduction,the pharmacological effect of a chemical compound relies on various interconnected properties of both the chemical molecule and the biological system [12]. Nonsteroidal Anti- Inflammatory drugs (NSAIDS) inhibit COX-1 and COX2 enzymes which through the pathway of prostaglandins modulate inflammation processes. This inhibition in cyclooxygenase enzymes leads to reduced inflammation but that does not come without cost: undesirable effects in heart and kidneys [18]. Thus, emerged the need for alternative drugs which would deal effectively with selectively COX-2-mediated inflammation but without the mentioned side effects. That led to the development of coxibs [32] which seemed to be very promising for their efficiency, however, it turned out in clinical trials that the possibility for an undesirable vascular event was big [5], having as an impact the withdrawal one of them (rofecoxib).

The chemical structure of a compound plays a paramount role as its interaction with biological targets will affect the biological and as a result, the pharmacological output.[14] COX-2 inhibitors can be divided in two substructural classes: tricyclics and not tricyclics. Focusing on the first class, those compounds are characterized by the presence of 1,2-diarylsubstitution on a central hetero or carbocyclic ring system with a characteristic methanesulfonyl, sulfonamido, azido, methanesulfonamide or pharmacophore-based tetrazole group on one of the aryl rings that plays a crucial role on COX-2 selectivity.



Rofecoxib
CS(=O)(=O)C1=CC=C(C=C1)C1=C(C(=O)OC1)C1=CC=CC=C1

Valdecoxib
CC1=C(C(=NO1)C1=CC=CC=C1)C1=CC=C(C=C1)S(N)(=O)=O

Figure 5.1: Structure and SMILES of Coxibs

Before having a closer a look at the suggested drugs, it is interesting to observe the number of drugs that each algorithm considered as the same. We can notice that the least number of drugs were suggested by topological fingerprint whereas the most by RDKit. Deploying ensemble methods increases the credibility of suggested drugs because the first one has neglected many similar drugs, whereas the other one has considered much more as such including more false positives as it was shown in the previous chapter.

Algorithm	Number of Similar Compounds
Topological Fingerprint	20
Morgan FIingerprint	377
RDKit Fingerprint	889
Maximum Common Subsequence	130

Table 5.1: Table containing LINGOS of Vanllin for q=4.

At table 5.2, are shown the similarity scores among coxibs and NSAIDS for the 4 different algorithms, verifying the chemical similarity thata characterizes them. In addition, at table 5.3, the scores that some of the suggested drugs -affecting other health systems- are provided. Those drugs have been proposed by at least one of the sugggested methods,equal weights, weighted and majority voting. As it becomes apparent, NSAIDS have greater scores for the different types of fingerprints and MCS, so it was easier to be recognised as similar to coxibs.

NSAIDs	Topological	Morgan	RDKit	MCS
Aceclofenac	0.225	0.476	0.523	0.289
Bromfenac	0.537	0.512	0.552	0.263
Etofenamate	0.233	0.513	0.568	0.375
Fenbufen	0.459	0.661	0.639	0.355
Fenoprofen	0.253	0.548	0.653	0.309
Flurbiprofen	0.393	0.585	0.509	0.333
Ketoprofen	0.333	0.66	0.486	0.38
Meloxicam	0.275	0.588	0.558	0.333
Naproxen	0.246	0.44	0.654	0.428
Nepafenac	0.523	0.643	0.528	0.333
Oxaprozin	0.493	0.714	0.489	0.315
Piroxicam	0.333	0.619	0.459	0.476
Sulindac	0.395	0.595	0.553	0.496
Suprofen	0.301	0.504	0.527	0.333
Tiaprofenic acid	0.388	0.536	0.552	0.35

Table 5.2: Similarity between NSAIDS and Coxibs

Having a look at table 5.4, it is expected the presence of NSAIDS drugs to the ones similar to coxibs. As it was mentioned in the introduction of this application, coxibs are a subgroup of NSAIDS, exclusively inhibiting COX-2 mediated inflammation. Apart from these, cardiovascular drugs are found as similar. Coxibs have adverse effects on cardiovascular system, so it makes sense drugs affecting the same organ system to be found with the same chemical properties. In addition, COX-2 has an antiviral effect as it has been shown that it can suppress Herpes simplex Virus type 1 reactivation [17]. An even deeper and analytical study has been carried out by Kohler et. al [25] who examined the impact of drugs affecting the circulatory system -mentioned above- through their toxicity and how they can be used as antivirals.One of those mentioned are coxibs.

Moving to central nervous system medications, COX-2 do have effect on this systems. It has been reported that Celecoxib has analgesic effect through endogenous opioid and cannabinoid mechanisms[43] and what is more, it is assessed their role in dealing with acute pain[42] and if they also have effect on diseases of this system like Alzheimer [15]. The subgroup of our

Drugs	Topological	Morgan	RDKit	MCS
Acetohexamide	0.513	0.487	0.618	0.5
Alverine	0.264	0.518	0.621	0.261
Amitriptyline	0.457	0.585	0.261	0.571
Armodafinil	0.125	0.655	0.589	0.309
Belinostat	0.472	0.612	0.59	0.476
Bromfenac	0.592	0.547	0.555	0.357
Butenafine	0.435	0.6	0.603	0.357
Chlorcyclizine	0.266	0.492	0.642	0.377
Dienestrol	0.4	0.627	0.523	0.357
Diethylstilbestrol	0.507	0.627	0.688	0.357
Diphenhydramine	0.205	0.531	0.5314	0.285
Mafenide	0.453	0.532	0.558	0.236
Modafinil	0.2	0.614	0.639	0.355
Ospemifene	0.591	0.609	0.601	0.38
Oxacillin	0.48	0.586	0.588	0.378
Protriptyline	0.314	0.595	0.552	0.238
Sulfacetamide	0.508	0.545	0.681	0.377
Tamoxifen	0.583	0.615	0.601	0.357
Tolbutamide	0.471	0.541	0.57	0.377
Toremifene	0.575	0.589	0.608	0.38

Table 5.3: Similarity between suggested drugs and Coxibs

interest has been found to increase gastrointestinal adverse outcome and this effect might be the reason for the association of the similarity with GI drugs.

Proving its potential as a great therapeutic target[52], aiming at multiple systems of the human body, COX-2 needs to be further explored and discover safer versions of coxibs or even better that have the same pharmacological effect by repurposing already approved drugs.

5.2 Model 2: An application in cancer through metabolic pathways.

Nowadays, one of the most prominent ways to cure cancer is by using chemotherapy. This treatment, though, has many disadvantages, with some of them being a great number of side-effects, cytotoxicity of the drugs and also, chemotherapy resistance. Having this impact, targeted therapies for cancer have received more attention. Metabolic pathways are networks including genes, proteins and metabolite reactions which interact with each other. All these are organized and function in an harmonic way but not in

Drug Category	Majority Voting	Weighting	Equal Weights
Anticancer	16	5	3
Anticonvulsant	7	-	-
Antihistamines	11	-	-
Antimicrobials	40	10	7
Antiviral	9	3	1
CNS	24	4	-
CVD	28	8	2
Diabetes (type II)	7	3	2
GI-disorders	5	-	-
Hormone related therapy	7	4	2
NSAIDs	12	13	9
Opioid analogues	5	-	-
Stimulants	6	-	-
Otherc	24	2	1
Total	201	52	27

Table 5.4: Drug categories of suggested repositioned drugs.

cancer cells. Their networks are dysregulated and proliferation is increased to a great degree by changing the metabolic pathways that this kind of cells use. This was reported by Otto Warburg, who stated that cancer cells metabolize glucose by glycolysis, which is a less efficient pathway compared to oxidative phosphorylation that normal cells use. At this point, a lot of things are altered regarding the way that malignant cells use their pathways in order to cover their increased energy needs [21]. As a result, finding drugs affecting metabolic pathways could halt the tumor genesis by affecting procedures like apoptosis and proliferation. This could be achieved by repositioning drugs and this was the goal of applying the proposed model.

Having created the dataset we get the drugs with ATC code 'L01' which were used in metabolic pathways, so these were the positive samples of the data. The negatives were anticancer drugs, not associating with targets in metabolic pathways and drugs either being associated with metabolic targets or not.

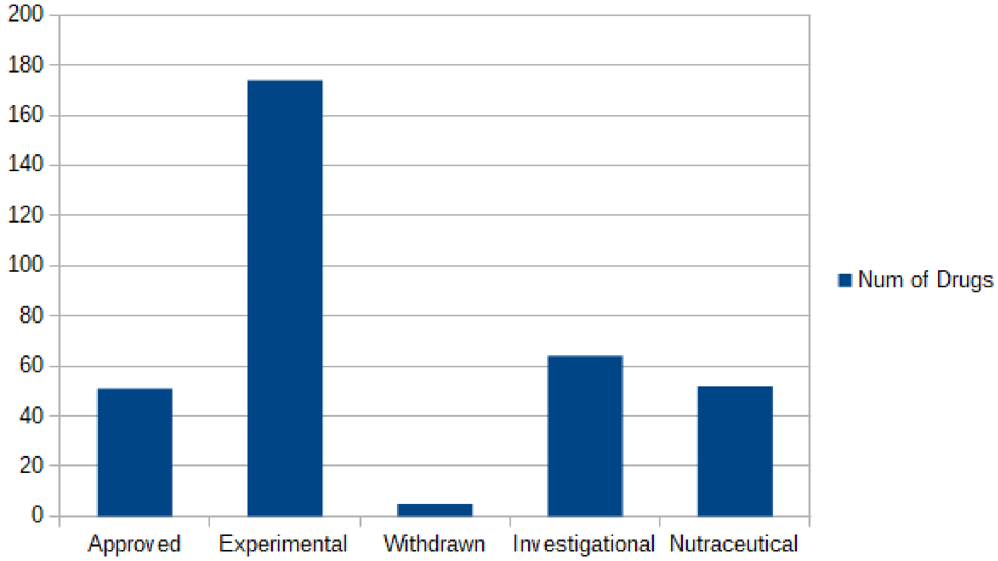
5.2.1 Drugs

The drugs that are found to be used as repositioned are 284 and for metabolic pathways are 63. Some of the drugs that are proposed by the model are demonstrated in the following table but the list with all drugs can be found in the appendix.

Drugs
Inosinic Acid
Ademetionine
Tegafur-uracil
Resveratrol
Didanosine
Guanosine
Floxuridine
Zebularine
Thymectacin

Table 5.5: Suggested repositioned drugs

The names of the drugs might seem a bit strange to the reader but this is because most of them are experimental as it can be seen in figure 5.2 .



[H]

Figure 5.2: Type of drugs

The number behind having so many experimental drugs lies on the fact that DrugBank can characterize an approved drug as an experimental or investigational if it has been so for other diseases. Therefore, it can simultaneously characterize a drug as approved, experimental and investigational. Some of the approved ones have been used to treat alternative diseases than those that were initially designed leading to the great number of experimental. An interesting result is the presence of so many nutraceutical drugs. That might not be the norm when thinking about curing a disease, especially cancer, however, many studies prove that such drugs could actually work. An example of this is Phloretin, a compound found in the leaves of apple trees and is used for colon cancer due to lowering the rates of glycolysis [41].

5.2.2 Drug Categories

One more step in analyzing the recommended drugs is to find out their pharmaceutical action and the diseases they are supposed to treat. This was achieved by using ATC code which categorized drugs and some of the results are demonstrated in the following table- the whole table is in the appendix.

ATC code	Category	Number of drugs
D06BB	Antivirals	2
G01AX	Other antiinfectives and antiseptics	2
J05AF	Nucleoside and nucleotide reverse transcriptase inhibitors	2
J05AB	Nucleosides and nucleotides excl. reverse transcriptase inhibitors	7
A11CA	Vitamin A, plain	9
D10AD	Retinooids for topical use in acne	9
R01AX	Other nasal preparations	9
V04CB	Tests for fat absorption	9
S01XA	Other ophalmologicals	11
A16AA	Amino Acids and Derivatives	13

Table 5.6: Drugs categorized according to their ATC

From this analysis and combined with scientific literature of the relevant field, it comes out that some of the drugs belong to categories with established repositioned drugs for cancer. In this categories belong antivirals (D06BB,J05AF, J05AB)and antiinfectives(G01AX) as shown by Sleire et.al[50]. As for vitamin A category, it is proved that these have similar structure or biological activity with retinooids which promote differentiation and cancer cell death [34]. Li et. al [30] tested amino acid derivatives(A16AA) for their antiproliferative activities against cancer and found one of them inducing apoptosis and prolonged cell cycle progression. All these indicate a possible use of these drugs as anticancers.

5.2.3 Targets

As it was mentioned before, targets are very important for metabolic pathways. Finding effective drugs for this problem means inhibiting or more generally speaking to affect a target, leading to an impact on the pathway. Analyzing the results we got the mostly affected targets, some of which are presented in figure 5.3. Each one of the repositioned drugs interacts with one or more targets, thus, affecting many of them simultaneously and this is the reason behind it.

Targets	Number of associations
Cytosolic purine 5′-nucleotidase	9
Galactose-1-phosphate uridylyltransferase	9
Adenosylhomocysteinase	9
UDP-glucose 4-epimerase	10
Glyceraldehyde-3-phosphate dehydrogenase	12
Citrate synthase, mitochondrial	12
Purine nucleoside phosphorylase	14
Thymidylate synthase	19

Table 5.7: Targets and number of associations with drugs

There are more of them presented in the appendix.

The figure below provides an example with target Dihydrofolate reductase which is related to three anticancer drugs (Methotreate, Pemetrexed and Pralaxerate) and also is related to two of the repositioned experimental drugs. As it was stated in the beginning of this thesis, targets with similar binding sites could bind similar ligands and this is an example of this.

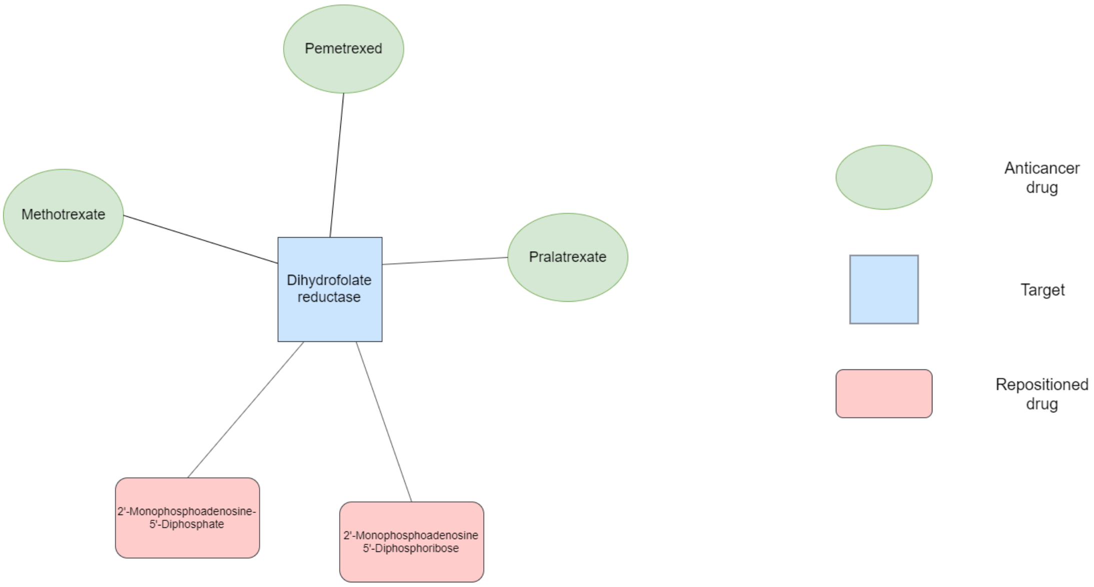


Figure 5.3: Anticancer drugs and repositioned interacting with target

5.2.4 Pathways

The next step is analyzing the pathways affected by the repositioned drugs. In the following table the pathways with the most drug-target associations are presented. The reader should keep in mind that a target might be present in more than one pathways.

Pathway Name	Number of drugs affecting it
Warburg Effect	6
Aspartate Metabolism	6
Arachidonic Acid Metabolism	7
Arginine and Proline Metabolism	7
Lactose Synthesis	7
Galactose Metabolism	7
Nicotinate and Nicotinamide Metabolism	7
Selenoamino Acid Metabolism	7
Methionine Metabolism	8
Retinol Metabolism	9
Purine Metabolism	19
Pyrimidine Metabolism	27

Table 5.8: Pathways and affected by repositioned drugs

Looking for their relationship with cancer, many of them were found to be related with it. Hence, it was needed to look for the types of pathways and what are their functions. KEGG Pathways and Biological Magnetic Resonance Data Bank are utilized in order to find the type of the pathways during their analysis. So, they are categorized in the main metabolic pathways types.

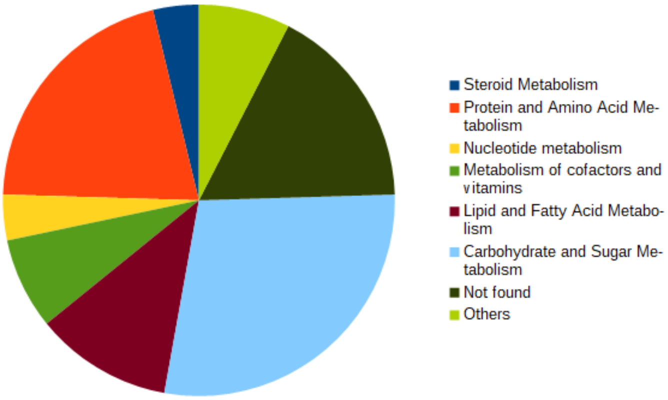


Figure 5.4: Categories of the pathways affected

Carbohydrate and sugar metabolism pathways are responsible for producing the necessary energy for the cell. As it has been shown by Warburg, cancer cells have increased needs for energy as a result of preferring the less effective glycolysis over oxidate phosphorylation. This leads to trigger other pathways to cover the energy demands of that kind of cell.[21] One more kind of pathways which has sparked the interest of scientists recently is fatty acid metabolic pathways. Genes involved in fatty acid synthesis or fatty acid oxidation have been correlated with tumor phenotypes like metastasis, therapeutic resistance and relapse [8] [28]. Regarding Protein and Amino Acid Metabolism, one of the largest groups of pathways , it has been shown that cancer cell metabolism is characterized by increased nitrogen demand, consumption of amino acids and upregulation of corresponding transporters, need for some specific nonessential amino acids and enzymes [31].

As a result, the suggested drugs of the model 2, seem to focus on pathways which are proved to be related with cancer, indicating that might actually affect them and have an impact as anticancer.

Chapter 6

Conclusions

Drug repositioning is a great field which can be assisted by chemical similarity. Methods such as fingerprints and MCS are valuable to take advantage of similar chemical structures as they provide all the necessary information. Utilizing similarity metrics and methods like voting and averaging the point of view of each different algorithm can contribute more to create a more general picture about chemical similarity of two drugs. Its application on coxibs delivered good and interesting results which worth validating using in vivo and in vitro experiments.

Regarding the second model, combining machine learning seems to perform excellent. Using Morgan fingerprints as a feature vector provides all the necessary information for the models to get trained and predict the outcome. Due to the small number of samples compared to the features, oversampling is necessary to apply. As it has been demonstrated, machine learning perform in a very satisfying way and when combined using the advantages of each one they can perform even better. The differences in performance among ensemble and other methods, like stacking or bagging were minor but all enhanced the outcome during cross validation and testing.

The results by applying Ensemble on DrugBank are promising after the analysis. The indications deriving from the pathways that are affected, its kind and their relation to cancer, the common targets among anticancer and repositioned drugs and their drug categories, are strong and suggest that they could actually be used. Of course, they should be experimentally tested and approved in order to use them for such an action.

In conclusion, taking all these into consideration, chemical similarity using SMILES can be a powerful tool for *in silico* approaches like this one. Machine Learning models and Ensemble can play a significant role in drug repositioning through their performance. Thus, there are more to be done in order to exploit and gain much more for them, achieving to deal with rare and life threatening diseases, improving in this way, people's lives.

Chapter 7

Future Work

There is still a lot that can be done and extend the work of this thesis. At first, more data can be used in order to find drugs that can be repositioned. Data including drug target interactions, side effects of drug which can be used to treat another disease and clinical data are just some of them that can be useful. Almost all of them have been used separately in order to discover new used for existing drugs but not so much in drug similarity and combining those different data categories. Therefore, it becomes inevitable to use more databases such as KEGG for genes, UNIPROT for targets, SMPDB for pathways, Disease Ontology for diseases and OMIM for side effects. Combining their data will assist to extract more knowledge and new associations.

In the computation part, what can be done for class imbalancing is to check more methods, such as under sampling and creating new artificial data. Regarding the machine learning part, it it would be interesting to try semi-supervised learning methods. To be more specific, having a few positive samples and trying to label much more unknown seems to be the case for that category. It is less used than supervised and unsupervised and there are only a few applications on biomedical projects, however, it might actually work. Apart from that, if the number of data increase and include categories as the ones suggested before, Deep Learning could also apply. There have already been some projects towards this direction and in our data-driven society, it will gain more and more ground.

In conclusion, those are just some suggestions for what could be done in the future. Drug repositioning, is an exciting field with great prospects for every passionate scientist, allowing to work in a wide variety of fields. Provided that the ambition and willingness exists so much more can be done and accomplished, expanding our knowledge about this field of study and hence, providing people with better healthcare and better life quality.

Chapter 8

Appendix

Table 8.1: Repositioned drugs

Repositioned Drugs
Adenosine-5-Diphosphoribose
Adenosine-5′-[Beta, Gamma-Methylene]Triphosphate
Guanosine-5′-Triphosphate
ATP
Phosphoaminophosphonic Acid-Adenylate Ester
Guanosine-5′-Diphosphate
Inosinic Acid
Guanosine-5′-Monophosphate
Ademetionine
3′-Oxo-Adenosine
D-Eritadenine
5′-S-ethyl-5′-thioadenosine
Adenosine-5′-Diphosphate-2′,3′-Vanadate
Adenosine 5′-phosphosulfate
S-adenosyl-L-homocysteine
′5′-O-(N-(L-Prolyl)-Sulfamoyl)Adenosine
5′-O-(N-(L-Cysteinyl)-Sulfamoyl)Adenosine
′5′-O-(N-(L-Alanyl)-Sulfamoyl)Adenosine
Tegafur-uracil
5-Fluorouridine
Adenosine monophosphate
2′-Monophosphoadenosine-5′-Diphosphate
Uridine-5′-Diphosphate
Uridine monophosphate

Continued on next page

Table 8.1 – continued from previous page

Repositioned Drugs
Phenyl-Uridine-5′-Diphosphate
Adenosine-3′-5′-Diphosphate
3′-Phosphate-Adenosine-5′-Phosphate Sulfate
Resveratrol
N6-ISOPENTENYL-ADENOSINE-5′-MONOPHOSPHATE
Nialamide
5-[2-(4-hydroxyphenyl)ethyl]benzene-1,3-diol
1-Phenylsulfonamide-3-Trifluoromethyl-5-Parabromophenylpyrazole
UP5
Didanosine
Guanosine
3-Deoxyguanosine
Inosine
GUANOSINE-2′,3′-O-ETHYLIDENEPHOSPHONATE
6-Chloropurine Riboside, 5′-Monophosphate
1-Deaza-Adenosine
Nebularine
Bis(Adenosine)-5′-Pentaphosphate
Tezacitabine
5′-Deoxy-5′-Methylthioadenosine
5′-Deoxy-5′-(Methylthio)-Tubercidin
Gamma-Arsono-Beta, Gamma-Methyleneadenosine-5′-Diphosphate
Zebularine
Thymidine-5′-Triphosphate
Floxuridine
Thymidine-5′-Phosphate
LY231514 Tetra Glu
2′-Deoxyuridine
LY341770
2′-5′dideoxyuridine
5-Fluoro-2′-Deoxyuridine-5′-Monophosphate
2′-Deoxycytidine-5′-Monophosphate
2′-deoxyuridylic acid
2′-Deoxyguanosine-5′-Monophosphate
Thymectacin
Continued on next page

Table 8.1 – continued from previous page

Repositioned Drugs
2-4-[2-(2-AMINO-4-OXO-4,7-DIHYDRO-3H-PYRROLO[2,3-D]PYRIMIDIN-5-YL)-ETHYL]-BENZOYLAMINO-3-METHYL-BUTYRIC ACID
Deoxyuridine-5′-Diphosphate
1-(2S,5S)-4-FLUORO-5-[(TRITYLOXY)METHYL]TETRAHYDROFURAN-2-YLPYRIMIDINE-2,4(1H,3H)-DIONE
Vitamin A
PICEATANNOL

Table 8.2: Pathways affected by suggestd drugs

Pathway	Appearances	Pathway Category
Glycolysis	1	Carbohydrate and Sugar Metabolism
Glucose-Alanine Cycle	1	Carbohydrate and Sugar Metabolism
Amino Sugar Metabolism	1	Carbohydrate and Sugar Metabolism
Starch and Sucrose Metabolism	1	Carbohydrate and Sugar Metabolism
Inositol Phosphate Metabolism	1	Lipid and Fatty Acid Metabolism
Inositol Metabolism	1	Lipid and Fatty Acid Metabolism
Glycerolipid Metabolism	1	Lipid and Fatty Acid Metabolism
Glycerol Phosphate Shuttle	1	Metabolism of cofactors and vitamins
Vitamin B6 Metabolism	1	Metabolism of other amino acids
Folate Metabolism	1	n/a
beta-Alanine Metabolism	1	n/a
Homocysteine Degradation	1	n/a
Degradation of Superoxides	1	n/a
Carnitine Synthesis	1	n/a
Catecholamine Biosynthesis	1	Protein and Amino Acid Metabolism

Continued on next page

Table 8.2 – continued from previous page

Target	Num of associations	
Glutathione Metabolism	1	Protein and Amino Acid Metabolism
Propanoate Metabolism	2	Metabolism of other amino acids
Pyruvate Metabolism	2	Alcohol Metabolism
Pentose Phosphate Pathway	2	Carbohydrate and Sugar Metabolism
Gluconeogenesis	2	Carbohydrate and Sugar Metabolism
Beta Oxidation of Very Long Chain Fatty Acids	2	Metabolism of cofactors and vitamins
Fatty Acid Metabolism	2	Metabolism of cofactors and vitamins
Oxidation of Branched-Chain Fatty Acids	2	Metabolism of cofactors and vitamins
Ammonia Recycling	2	n/a
Spermidine and Spermine Biosynthesis	2	Protein and Amino Acid Metabolism
Phenylalanine and Tyrosine Metabolism	2	Protein and Amino Acid Metabolism
Mitochondrial Beta-Oxidation of Long Chain Saturated Fatty Acids	2	Steroid Metabolism
Pterine Biosynthesis	3	n/a
Histidine Metabolism	3	Protein and Amino Acid Metabolism
Urea Cycle	3	Protein and Amino Acid Metabolism
Androgen and Estrogen Metabolism	3	Protein and Amino Acid Metabolism
Betaine Metabolism	4	Carbohydrate and Sugar Metabolism
Ethanol Degradation	4	Carbohydrate and Sugar Metabolism
Glutamate Metabolism	4	Carbohydrate and Sugar Metabolism
Sulfate/Sulfite Metabolism	4	Carbohydrate and Sugar Metabolism
Fructose and Mannose Degradation	4	Lipid and Fatty Acid Metabolism

Continued on next page

Table 8.2 – continued from previous page

Target	Num of associations	
Tyrosine Metabolism	4	Protein and Amino Acid Metabolism
Estrone Metabolism	4	Steroid Metabolism
Glycine and Serine Metabolism	5	Carbohydrate and Sugar Metabolism
Nucleotide Sugars Metabolism	5	Lipid and Fatty Acid Metabolism
Mitochondrial Electron Transport Chain	5	Nucleotide metabolism
Warburg Effect	6	Nucleotide metabolism
Aspartate Metabolism	6	Protein and Amino Acid Metabolism
Arachidonic Acid Metabolism	7	Carbohydrate and Sugar Metabolism
Arginine and Proline Metabolism	7	Carbohydrate and Sugar Metabolism
Lactose Synthesis	7	Carbohydrate and Sugar Metabolism
Galactose Metabolism	7	Lipid and Fatty Acid Metabolism
Nicotinate and Nicotinamide Metabolism	7	n/a
Selenoamino Acid Metabolism	7	Peptide Hormone Metabolism
Methionine Metabolism	8	Carbohydrate and Sugar Metabolism
Retinol Metabolism	9	n/a
Purine Metabolism	19	Protein and Amino Acid Metabolism
Pyrimidine Metabolism	27	Protein and Amino Acid Metabolism

Table 8.3: Known anticancer and suggested drugs sharing common targets

Suggested Drug	Target	Known Drug
ATP	Abelson tyrosine-protein kinase 2	Dasatinib
ATP	Multidrug resistance protein 1	Celecoxib

Continued on next page

Table 8.3 – continued from previous page

Target	Num of associations	
ATP	Tyrosine-protein kinase ABL1	Dasatinib
ATP	Cystic fibrosis transmembrane conductance regulator	Lonidamine
Phosphoaminophosphonic Acid-Adenylate Ester	Cystic fibrosis transmembrane conductance regulator	Lonidamine
Phosphoaminophosphonic Acid-Adenylate Ester	Hexokinase-1	Lonidamine
Phosphoaminophosphonic Acid-Adenylate Ester	Tyrosine-protein kinase Lck	Dasatinib
Phosphoaminophosphonic Acid-Adenylate Ester	Ephrin type-A receptor 2	Dasatinib
Guanosine-5′-Monophosphate	Bifunctional purine biosynthesis protein PURH	Pemetrexed
Guanosine-5′-Monophosphate	Amidophosphoribosyltransferase	Dasatinib
Guanosine-5′-Monophosphate	Amidophosphoribosyltransferase	Dasatinib
Tegafur-uracil	Thymidylate synthase	Raltitrexed
Tegafur-uracil	Thymidylate synthase	Trifluridine
Tegafur-uracil	Thymidylate synthase	Gemcitabine
Tegafur-uracil	Thymidylate synthase	Pemetrexed
Tegafur-uracil	Thymidylate synthase	Capecitabine
Tegafur-uracil	Thymidylate synthase	Pralatrexate
Tegafur-uracil	Thymidylate synthase	Tegafur
Tegafur-uracil	Thymidylate synthase	Tegafur
Uridine monophosphate	Thymidylate synthase	Raltitrexed
Uridine monophosphate	Thymidylate synthase	Trifluridine
Uridine monophosphate	Thymidylate synthase	Gemcitabine
Uridine monophosphate	Thymidylate synthase	Pemetrexed
Uridine monophosphate	Thymidylate synthase	Capecitabine
Uridine monophosphate	Thymidylate synthase	Pralatrexate
Uridine monophosphate	Thymidylate synthase	Tegafur
Uridine monophosphate	Thymidylate synthase	Tegafur
Floxuridine	Thymidylate synthase	Raltitrexed
Floxuridine	Thymidylate synthase	Trifluridine
Floxuridine	Thymidylate synthase	Gemcitabine
Floxuridine	Thymidylate synthase	Pemetrexed
Floxuridine	Thymidylate synthase	Capecitabine
Floxuridine	Thymidylate synthase	Pralatrexate
Floxuridine	Thymidylate synthase	Tegafur
Floxuridine	Thymidylate synthase	Tegafur

Continued on next page

Table 8.3 – continued from previous page

Target	Num of associations	
Thymidine-5′-Phosphate	Thymidylate synthase	Raltitrexed
Thymidine-5′-Phosphate	Thymidylate synthase	Trifluridine
Thymidine-5′-Phosphate	Thymidylate synthase	Gemcitabine
Thymidine-5′-Phosphate	Thymidylate synthase	Pemetrexed
Thymidine-5′-Phosphate	Thymidylate synthase	Capecitabine
Thymidine-5′-Phosphate	Thymidylate synthase	Pralatrexate
Thymidine-5′-Phosphate	Thymidylate synthase	Tegafur
Thymidine-5′-Phosphate	Thymidylate synthase	Tegafur
Thymidine-5′-Phosphate	Thymidylate synthase	Raltitrexed
Thymidine-5′-Phosphate	Thymidylate synthase	Trifluridine
Thymidine-5′-Phosphate	Thymidylate synthase	Gemcitabine
Thymidine-5′-Phosphate	Thymidylate synthase	Pemetrexed
Thymidine-5′-Phosphate	Thymidylate synthase	Capecitabine
Thymidine-5′-Phosphate	Thymidylate synthase	Pralatrexate
Thymidine-5′-Phosphate	Thymidylate synthase	Tegafur
Thymidine-5′-Phosphate	Thymidylate synthase	Tegafur
LY231514 Tetra Glu	Thymidylate synthase	Raltitrexed
LY231514 Tetra Glu	Thymidylate synthase	Trifluridine
LY231514 Tetra Glu	Thymidylate synthase	Gemcitabine
LY231514 Tetra Glu	Thymidylate synthase	Pemetrexed
LY231514 Tetra Glu	Thymidylate synthase	Capecitabine
LY231514 Tetra Glu	Thymidylate synthase	Pralatrexate
LY231514 Tetra Glu	Thymidylate synthase	Tegafur
LY231514 Tetra Glu	Thymidylate synthase	Tegafur
2′-Deoxyuridine	Thymidylate synthase	Raltitrexed
2′-Deoxyuridine	Thymidylate synthase	Trifluridine
2′-Deoxyuridine	Thymidylate synthase	Gemcitabine
2′-Deoxyuridine	Thymidylate synthase	Pemetrexed
2′-Deoxyuridine	Thymidylate synthase	Capecitabine
2′-Deoxyuridine	Thymidylate synthase	Pralatrexate
2′-Deoxyuridine	Thymidylate synthase	Tegafur
2′-Deoxyuridine	Thymidylate synthase	Tegafur
LY341770	Thymidylate synthase	Raltitrexed
LY341770	Thymidylate synthase	Trifluridine
LY341770	Thymidylate synthase	Gemcitabine
LY341770	Thymidylate synthase	Pemetrexed
LY341770	Thymidylate synthase	Capecitabine
LY341770	Thymidylate synthase	Pralatrexate
LY341770	Thymidylate synthase	Tegafur

Continued on next page

Table 8.3 – continued from previous page

Target	Num of associations	
LY341770	Thymidylate synthase	Tegafur
2′-5′dideoxyuridine	Thymidylate synthase	Raltitrexed
2′-5′dideoxyuridine	Thymidylate synthase	Trifluridine
2′-5′dideoxyuridine	Thymidylate synthase	Gemcitabine
2′-5′dideoxyuridine	Thymidylate synthase	Pemetrexed
2′-5′dideoxyuridine	Thymidylate synthase	Capecitabine
2′-5′dideoxyuridine	Thymidylate synthase	Pralatrexate
2′-5′dideoxyuridine	Thymidylate synthase	Tegafur
2′-5′dideoxyuridine	Thymidylate synthase	Tegafur
5-Fluoro-2′-Deoxyuridine-5′-Monophosphate	Thymidylate synthase	Raltitrexed
5-Fluoro-2′-Deoxyuridine-5′-Monophosphate	Thymidylate synthase	Trifluridine
5-Fluoro-2′-Deoxyuridine-5′-Monophosphate	Thymidylate synthase	Gemcitabine
5-Fluoro-2′-Deoxyuridine-5′-Monophosphate	Thymidylate synthase	Pemetrexed
5-Fluoro-2′-Deoxyuridine-5′-Monophosphate	Thymidylate synthase	Capecitabine
5-Fluoro-2′-Deoxyuridine-5′-Monophosphate	Thymidylate synthase	Pralatrexate
5-Fluoro-2′-Deoxyuridine-5′-Monophosphate	Thymidylate synthase	Tegafur
5-Fluoro-2′-Deoxyuridine-5′-Monophosphate	Thymidylate synthase	Tegafur
5-Fluoro-2′-Deoxyuridine-5′-Monophosphate	Thymidylate synthase	Raltitrexed
5-Fluoro-2′-Deoxyuridine-5′-Monophosphate	Thymidylate synthase	Trifluridine
5-Fluoro-2′-Deoxyuridine-5′-Monophosphate	Thymidylate synthase	Gemcitabine
5-Fluoro-2′-Deoxyuridine-5′-Monophosphate	Thymidylate synthase	Pemetrexed
5-Fluoro-2′-Deoxyuridine-5′-Monophosphate	Thymidylate synthase	Capecitabine
5-Fluoro-2′-Deoxyuridine-5′-Monophosphate	Thymidylate synthase	Pralatrexate
5-Fluoro-2′-Deoxyuridine-5′-Monophosphate	Thymidylate synthase	Tegafur

Continued on next page

Table 8.3 – continued from previous page

Target	Num of associations	
5-Fluoro-2′-Deoxyuridine-5′-MonophosphateM	Thymidylate synthase	Tegafur
2′-Deoxycytidine-5′-Monophosphate	Thymidylate synthase	Raltitrexed
2′-Deoxycytidine-5′-Monophosphate	Thymidylate synthase	Trifluridine
2′-Deoxycytidine-5′-Monophosphate	Thymidylate synthase	Gemcitabine
2′-Deoxycytidine-5′-Monophosphate	Thymidylate synthase	Pemetrexed
2′-Deoxycytidine-5′-Monophosphate	Thymidylate synthase	Capecitabine
2′-Deoxycytidine-5′-Monophosphate	Thymidylate synthase	Pralatrexate
2′-Deoxycytidine-5′-Monophosphate	Thymidylate synthase	Tegafur
2′-Deoxycytidine-5′-Monophosphate	Thymidylate synthase	Tegafur
2′-deoxyuridylic acid	Thymidylate synthase	Raltitrexed
2′-deoxyuridylic acid	Thymidylate synthase	Trifluridine
2′-deoxyuridylic acid	Thymidylate synthase	Gemcitabine
2′-deoxyuridylic acid	Thymidylate synthase	Pemetrexed
2′-deoxyuridylic acid	Thymidylate synthase	Capecitabine
2′-deoxyuridylic acid	Thymidylate synthase	Pralatrexate
2′-deoxyuridylic acid	Thymidylate synthase	Tegafur
2′-deoxyuridylic acid	Thymidylate synthase	Tegafur
2′-deoxyuridylic acid	Thymidylate synthase	Raltitrexed
2′-deoxyuridylic acid	Thymidylate synthase	Trifluridine
2′-deoxyuridylic acid	Thymidylate synthase	Gemcitabine
2′-deoxyuridylic acid	Thymidylate synthase	Pemetrexed
2′-deoxyuridylic acid	Thymidylate synthase	Capecitabine
2′-deoxyuridylic acid	Thymidylate synthase	Pralatrexate
2′-deoxyuridylic acid	Thymidylate synthase	Tegafur
2′-deoxyuridylic acid	Thymidylate synthase	Tegafur
2′-deoxyuridylic acid	Thymidylate synthase	Raltitrexed
2′-deoxyuridylic acid	Thymidylate synthase	Trifluridine
2′-deoxyuridylic acid	Thymidylate synthase	Gemcitabine
2′-deoxyuridylic acid	Thymidylate synthase	Pemetrexed
2′-deoxyuridylic acid	Thymidylate synthase	Capecitabine

Continued on next page

Table 8.3 – continued from previous page

Target	Num of associations	
2′-deoxyuridylic acid	Thymidylate synthase	Pralatrexate
2′-deoxyuridylic acid	Thymidylate synthase	Tegafur
2′-deoxyuridylic acid	Thymidylate synthase	Tegafur
2′-deoxyuridylic acid	Thymidylate synthase	Raltitrexed
2′-deoxyuridylic acid	Thymidylate synthase	Trifluridine
2′-deoxyuridylic acid	Thymidylate synthase	Gemcitabine
2′-deoxyuridylic acid	Thymidylate synthase	Pemetrexed
2′-deoxyuridylic acid	Thymidylate synthase	Capecitabine
2′-deoxyuridylic acid	Thymidylate synthase	Pralatrexate
2′-deoxyuridylic acid	Thymidylate synthase	Tegafur
2′-deoxyuridylic acid	Thymidylate synthase	Tegafur
2′-Deoxyguanosine-5′-Monophosphate	Thymidylate synthase	Raltitrexed
2′-Deoxyguanosine-5′-Monophosphate	Thymidylate synthase	Trifluridine
2′-Deoxyguanosine-5′-Monophosphate	Thymidylate synthase	Gemcitabine
2′-Deoxyguanosine-5′-Monophosphate	Thymidylate synthase	Pemetrexed
2′-Deoxyguanosine-5′-Monophosphate	Thymidylate synthase	Capecitabine
2′-Deoxyguanosine-5′-Monophosphate	Thymidylate synthase	Pralatrexate
2′-Deoxyguanosine-5′-Monophosphate	Thymidylate synthase	Tegafur
2′-Deoxyguanosine-5′-Monophosphate	Thymidylate synthase	Tegafur
Thymectacin	Thymidylate synthase	Raltitrexed
Thymectacin	Thymidylate synthase	Trifluridine
Thymectacin	Thymidylate synthase	Gemcitabine
Thymectacin	Thymidylate synthase	Pemetrexed
Thymectacin	Thymidylate synthase	Capecitabine
Thymectacin	Thymidylate synthase	Pralatrexate
Thymectacin	Thymidylate synthase	Tegafur
Thymectacin	Thymidylate synthase	Tegafur

Continued on next page

Table 8.3 – continued from previous page

Target	Num of associations	
2-4-[2-(2-AMINO-4-OXO-4,7-DIHYDRO-3H-PYRROLO[2,3-D]PYRIMIDIN-5-YL)-ETHYL]-BENZOYLAMINO-3-METHYL-BUTYRIC ACID	Thymidylate synthase	Raltitrexed
2-4-[2-(2-AMINO-4-OXO-4,7-DIHYDRO-3H-PYRROLO[2,3-D]PYRIMIDIN-5-YL)-ETHYL]-BENZOYLAMINO-3-METHYL-BUTYRIC ACID	Thymidylate synthase	Trifluridine
2-4-[2-(2-AMINO-4-OXO-4,7-DIHYDRO-3H-PYRROLO[2,3-D]PYRIMIDIN-5-YL)-ETHYL]-BENZOYLAMINO-3-METHYL-BUTYRIC ACID	Thymidylate synthase	Gemcitabine
2-4-[2-(2-AMINO-4-OXO-4,7-DIHYDRO-3H-PYRROLO[2,3-D]PYRIMIDIN-5-YL)-ETHYL]-BENZOYLAMINO-3-METHYL-BUTYRIC ACID	Thymidylate synthase	Pemetrexed
2-4-[2-(2-AMINO-4-OXO-4,7-DIHYDRO-3H-PYRROLO[2,3-D]PYRIMIDIN-5-YL)-ETHYL]-BENZOYLAMINO-3-METHYL-BUTYRIC ACID	Thymidylate synthase	Capecitabine
2-4-[2-(2-AMINO-4-OXO-4,7-DIHYDRO-3H-PYRROLO[2,3-D]PYRIMIDIN-5-YL)-ETHYL]-BENZOYLAMINO-3-METHYL-BUTYRIC ACID	Thymidylate synthase	Pralatrexate
2-4-[2-(2-AMINO-4-OXO-4,7-DIHYDRO-3H-PYRROLO[2,3-D]PYRIMIDIN-5-YL)-ETHYL]-BENZOYLAMINO-3-METHYL-BUTYRIC ACID	Thymidylate synthase	Tegafur
2-4-[2-(2-AMINO-4-OXO-4,7-DIHYDRO-3H-PYRROLO[2,3-D]PYRIMIDIN-5-YL)-ETHYL]-BENZOYLAMINO-3-METHYL-BUTYRIC ACID	Thymidylate synthase	Tegafur

Continued on next page

Table 8.3 – continued from previous page

Target	Num of associations	
2′-Monophosphoadenosine-5′-Diphosphate	Dihydrofolate reductase	Methotrexate
2′-Monophosphoadenosine-5′-Diphosphate	Dihydrofolate reductase	Pemetrexed
2′-Monophosphoadenosine-5′-Diphosphate	Dihydrofolate reductase	Pralatrexate
Uridine-5′-Diphosphate	UMP-CMP kinase	Gemcitabine
UP5	UMP-CMP kinase	Gemcitabine
Resveratrol	Prostaglandin G/H synthase 2	Celecoxib
1-Phenylsulfonamide-3-Trifluoromethyl-5-Parabromophenylpyrazole	Prostaglandin G/H synthase 2	Celecoxib
Resveratrol	Arachidonate 5-lipoxygenase	Masoprocol
Resveratrol	Estrogen receptor	Mitotane
Nialamide	Amine oxidase [flavin-containing] B	Procarbazine
Nialamide	Amine oxidase [flavin-containing] A	Procarbazine
Didanosine	Purine nucleoside phosphorylase	Cladribine
Guanosine	Purine nucleoside phosphorylase	Cladribine
Guanosine	Purine nucleoside phosphorylase	Cladribine
3-Deoxyguanosine	Purine nucleoside phosphorylase	Cladribine
Inosine	Purine nucleoside phosphorylase	Cladribine
GUANOSINE-2′,3′-O-ETHYLIDENEPHOSPHONATE	Purine nucleoside phosphorylase	Cladribine
1-Deaza-Adenosine	Adenosine deaminase	Pentostatin
Nebularine	Adenosine deaminase	Pentostatin
Tezacitabine	Ribonucleoside-diphosphate reductase large subunit	Cladribine
Tezacitabine	Ribonucleoside-diphosphate reductase large subunit	Gemcitabine
Tezacitabine	Ribonucleoside-diphosphate reductase large subunit	Clofarabine
Tezacitabine	Ribonucleoside-diphosphate reductase large subunit	Fludarabine

Continued on next page

Table 8.3 – continued from previous page

Target	Num of associations	
Vitamin A	Retinal dehydrogenase 2	Tretinoin
Vitamin A	Retinal dehydrogenase 1	Tretinoin
Vitamin A	Hematopoietic prostaglandin D synthase	Tretinoin

Table 8.4: Targets interacting with repositioned drugs.

Target	Num of associations
6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase 1	1
GTP cyclohydrolase 1	1
Glycerol kinase	1
Glycogen phosphorylase, liver form	1
Guanylate kinase	1
Histamine N-methyltransferase	1
Histone-lysine N-methyltransferase SETD7	1
Hypoxanthine-guanine phosphoribosyltransferase	1
Inosine-5′-monophosphate dehydrogenase 1	1
Lecithin retinol acyltransferase	1
Leukotriene A-4 hydrolase	1
Nucleoside diphosphate kinase B	1
Prostaglandin G/H synthase 1	1
GMP reductase 1	1
Protein arginine N-methyltransferase 3	1
Retinal dehydrogenase 1	1
Retinal dehydrogenase 2	1
Retinol dehydrogenase 11	1
Retinol dehydrogenase 12	1
Retinol dehydrogenase 8	1
Ribokinase	1
Ribonucleoside-diphosphate reductase large subunit	1
Short-chain dehydrogenase/reductase 3	1
Solute carrier family 2, facilitated glucose transporter member 1	1
Sulfotransferase 1A1	1
Superoxide dismutase [Cu-Zn]	1
Thymidine kinase, cytosolic	1
Pyridoxal kinase	1
GDP-mannose 4,6 dehydratase	1

Continued on next page

Table 8.4 – continued from previous page

Target	Num of associations
cAMP-specific 3′,5′-cyclic phosphodiesterase 4D	1
Bifunctional purine biosynthesis protein PURH	1
ADP/ATP translocase 1	1
Dehydrogenase/reductase SDR family member 4	1
ATP synthase subunit alpha, mitochondrial	1
Arachidonate 5-lipoxygenase	1
ATP synthase subunit beta, mitochondrial	1
ATP synthase subunit gamma, mitochondrial	1
Arachidonate 15-lipoxygenase	1
Cytidine deaminase	1
All-trans-retinol 13,14-reductase	1
Adenylate kinase isoenzyme 1	1
Dihydropyrimidine dehydrogenase [NADP(+)]	2
Dihydrofolate reductase	2
Prostaglandin G/H synthase 2	2
S-adenosylmethionine decarboxylase proenzyme	2
Phenylethanolamine N-methyltransferase	2
S-methyl-5′-thioadenosine phosphorylase	2
Sulfotransferase family cytosolic 2B member 1	2
Tyrosine–tRNA ligase, cytoplasmic	2
UMP-CMP kinase	2
Adenosine deaminase	2
Asparagine synthetase [glutamine-hydrolyzing]	2
Phosphoglycerate kinase 1	2
Inositol-trisphosphate 3-kinase A	2
Deoxyuridine 5′-triphosphate nucleotidohydrolase, mitochondrial	2
Galactokinase	2
Glutathione synthetase	2
Beta-1,4-galactosyltransferase 1	2
Guanidinoacetate N-methyltransferase	2
Uridine-cytidine kinase-like 1	2
Bifunctional 3′-phosphoadenosine 5′-phosphosulfate synthase 1	2
Carbonyl reductase [NADPH] 1	2
Glycine N-methyltransferase	2
Adenylosuccinate synthetase isozyme 1	3
Fructose-1,6-bisphosphatase 1	3
Cytosolic purine 5′-nucleotidase	3
Bifunctional glutamate/proline–tRNA ligase	3
Amine oxidase [flavin-containing] A	3

Continued on next page

Table 8.4 – continued from previous page

Target	Num of associations
Hexokinase-1	3
Amidophosphoribosyltransferase	4
UDP-glucose 4-epimerase	4
Acetyl-coenzyme A synthetase 2-like, mitochondrial	4
Acetyl-coenzyme A synthetase, cytoplasmic	4
Cystathionine beta-synthase	4
S-adenosylmethionine synthase isoform type-2	4
Catechol O-methyltransferase	4
Glyceraldehyde-3-phosphate dehydrogenase	6
Galactose-1-phosphate uridylyltransferase	6
Argininosuccinate synthase	6
Estrogen sulfotransferase	6
Glutamate dehydrogenase 1, mitochondrial	7
Long-chain-fatty-acid–CoA ligase 1	8
Adenosylhomocysteinase	9
Purine nucleoside phosphorylase	12
Thymidylate synthase	19

Bibliography

[1] Ted T. Ashburn and Karl B. Thor. Drug repositioning: identifying and developing new uses for existing drugs. *Nature Reviews Drug Discovery*, 3(8):673–683, aug 2004.

[2] Leo Breiman. Bagging Predictors. pages 123—-140, 1994.

[3] Leo Breiman. Stacked regressions. *Machine Learning*, 24(1):49–64, jul 1996.

[4] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, oct 2001.

[5] C. P. Cannon and P. J. Cannon. COX-2 Inhibitors and Cardiovascular Risk. *Science*, 336(6087):1386–1387, jun 2012.

[6] Adrià Cereto-Massagué, María José Ojeda, Cristina Valls, Miquel Mulero, Santiago Garcia-Vallvé, and Gerard Pujadas. Molecular fingerprint similarity search in virtual screening. *Methods*, 71(C):58–63, 2015.

[7] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. Technical report, Department of Computer Science and Engineering, ENB 118 University of South Florida 4202, 2002.

[8] Ming Chen and Jiaoti Huang. The expanded role of fatty acid metabolism in cancer: new aspects and targets. *Precision Clinical Medicine*, 2(3):183–191, oct 2019.

[9] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, sep 1995.

[10] T. M. Cover and P. E. Hart. Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

[11] Lee R. Dice. Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26(3):297–302, jul 1945.

[12] Sean Ekins, Jair Lage de Siqueira-Neto, Laura-Isobel McCall, Malabika Sarker, Maneesh Yadav, Elizabeth L. Ponder, E. Adam Kallel, Danielle Kellar, Steven Chen, Michelle Arkin, Barry A. Bunin, James H. McKerrow, and Carolyn Talcott. Machine Learning Models and Pathway Genome Data Base for Trypanosoma cruzi Drug Discovery. *PLOS Neglected Tropical Diseases*, 9(6):e0003878, jun 2015.

[13] Jiashi Feng, Huan Xu, Shie Mannor, and Shuicheng Yan. Robust Logistic Regression and Classification. Technical report, EECS Department & ICSI UC Berkeley.

[14] Artur J. Ferreira and Mário A. T. Figueiredo. Boosting Algorithms: A Review of Methods, Theory, and Applications. In *Ensemble Machine Learning*, pages 35–85. Springer US, Boston, MA, 2012.

[15] Omidreza Firuzi and Domenico Praticò. Coxibs and Alzheimer's disease: Should they stay or should they go? *Annals of Neurology*, 59(2):219–228, feb 2006.

[16] Yoav Freund and Robert E Schapire. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55(1):119–139, aug 1997.

[17] Bryan M. Gebhardt, Emily D. Varnell, and Herbert E. Kaufman. Inhibition of cyclooxygenase 2 synthesis suppresses Herpes simplex virus type 1 reactivation. *Journal of Ocular Pharmacology and Therapeutics*, 21(2):114–120, apr 2005.

[18] Sam Harirforoosh, Waheed Asghar, and Fakhreddin Jamali. Adverse Effects of Nonsteroidal Antiinflammatory Drugs: An Update of Gastrointestinal, Cardiovascular and Renal Complications. *Journal of Pharmacy & Pharmaceutical Sciences*, 16(5):821, jan 2014.

[19] Kathrin Heikamp and Jürgen Bajorath. Large-scale similarity search profiling of ChEMBL compound data sets. *Journal of Chemical Information and Modeling*, 51(8):1831–1839, 2011.

[20] Mohammad Hossin and Nasir Sulaiman. A Review on Evaluation Metrics For Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5(2):1–11, 2015.

[21] Peggy P. Hsu and David M. Sabatini. Cancer cell metabolism: Warburg and beyond. *Cell*, 134(5):703–707, 2008.

[22] Robert W. Irving and Campbell B. Fraser. Maximal common subsequences and minimal common supersequences. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 807 LNCS, pages 173–183. Springer Verlag, 1994.

[23] Michael J. Keiser, Vincent Setola, John J. Irwin, Christian Laggner, Atheir I. Abbas, Sandra J. Hufeisen, Niels H. Jensen, Michael B. Kuijer, Roberto C. Matos, Thuy B. Tran, Ryan Whaley, Richard A. Glennon, Jérôme Hert, Kelan L. H. Thomas, Douglas D. Edwards, Brian K. Shoichet, and Bryan L. Roth. Predicting new molecular targets for known drugs. *Nature*, 462(7270):175–181, nov 2009.

[24] Ron Kohavi and Foster Provost. Glossary of terms. *Machine Learning*, 30(2):271–274, Feb 1998.

[25] James J. Kohler and William Lewis. Cardiovascular Toxicities of Life-Saving Drugs: Antiviral Therapy. In *Cardiotoxicity of Non-Cardiovascular Drugs*, pages 313–332. John Wiley and Sons, apr 2010.

[26] S. B. Kotsiantis. Decision trees: A recent overview, apr 2013.

[27] Sotiris B. Kotsiantis. Bagging and boosting variants for handling classifications problems: a survey. *The Knowledge Engineering Review*, 29(1):78–100, jan 2014.

[28] Ching Ying Kuo and David K. Ann. When fats commit crimes: Fatty acid metabolism, cancer stemness and therapeutic resistance, jul 2018.

[29] Kevin Lang, Yahoo Research, Edo Liberty, and Konstantin Shmakov. Stratified Sampling Meets Machine Learning. Technical report, 2016.

[30] Yang Li, Qizhi Zhang, Jun He, Wenmei Yu, Jie Xiao, Yu Guo, Xiaoming Zhu, and Yunmei Liu. Synthesis and biological evaluation of amino acid derivatives containing chrysin that induce apoptosis. *Natural Product Research*, pages 1–10, mar 2019.

[31] Michael J. Lukey, William P. Katt, and Richard A. Cerione. Targeting amino acid metabolism for cancer therapy, may 2017.

[32] L J Marnett and A S Kalgutkar. Cyclooxygenase 2 inhibitors: discovery, selectivity and the future. *Trends in pharmacological sciences*, 20(11):465–9, nov 1999.

[33] H. L. Morgan. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation*, 5(2):107–113, may 1965.

[34] L Nagy, V A Thomázy, G L Shipley, L Fésüs, W Lamph, R A Heyman, R A Chandraratna, and P J Davies. Activation of retinoid X receptors induces apoptosis in HL-60 cell lines. *Molecular and cellular biology*, 15(7):3540–51, jul 1995.

[35] Ramaswamy Nilakantan, Norman Bauman, J. Scott Dixon, and R. Venkataraghavan. Topological torsion: a new molecular descriptor for SAR applications. Comparison with other descriptors. *Journal of Chemical Information and Modeling*, 27(2):82–85, may 1987.

[36] Tudor I Oprea, Elebeoba E May, Andrei Leitão, and Alexander Tropsha. Computational systems chemical biology. *Methods in molecular biology (Clifton, N.J.)*, 672:459–88, 2011.

[37] Hakime Öztürk, Elif Ozkirimli, and Arzucan Özgür. A comparative study of SMILES-based compound similarity functions for drug-target interaction prediction. *BMC Bioinformatics*, 17(1):1–11, 2016.

[38] Arti Patle and Deepak Singh Chouhan. SVM kernel functions for classification. In *2013 International Conference on Advances in Technology and Engineering, ICATE 2013*, 2013.

[39] Nolen Joy Perualila-Tan, Ziv Shkedy, Willem Talloen, Hinrich W. H. Göhlmann, Marijke Van Moerbeke, and Adetayo Kasim. Weighted similarity-based clustering of chemical structures and bioactivity data in early drug discovery. *Journal of Bioinformatics and Computational Biology*, 14(04):1650018, 2016.

[40] Sudeep Pushpakom, Francesco Iorio, Patrick A Eyers, K Jane Escott, Shirley Hopper, Andrew Wells, Andrew Doig, Tim Guilliams, Joanna Latimer, Christine McNamee, Alan Norris, Philippe Sanseau, David Cavalla, and Munir Pirmohamed. Drug repurposing: progress, challenges and recommendations. 2018.

[41] Mahbuba Rahman and Mohammad Hasan. Cancer Metabolism and Drug Resistance. *Metabolites*, 5(4):571–600, sep 2015.

[42] Scott S Reuben. Update on the role of nonsteroidal anti-inflammatory drugs and coxibs in the management of acute pain. *Current Opinion in Anaesthesiology*, 20(5):440–450, oct 2007.

[43] R.M. Rezende, P. Paiva-Lima, W.G.P. Dos Reis, V.M. Camêlo, A. Faraco, Y.S. Bakhle, and J.N. Francischi. Endogenous Opioid and Cannabinoid Mechanisms Are Involved in the Analgesic Effects of Celecoxib in the Central Nervous System. *Pharmacology*, 89(3-4):127–136, 2012.

[44] Sereina Riniker and Gregory A Landrum. Similarity maps - a visualization strategy for molecular fingerprints and machine-learning methods. *Journal of Cheminformatics*, 5(1):43, dec 2013.

[45] Robert E. Schapire. Explaining AdaBoost. In *Empirical Inference*, pages 37–52. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[46] Konstantinos Koutroumbas Sergios Theodoridis. Classifiers Based on Bayes Decision Theory. Technical report, 2008.

[47] Robert P. Sheridan and Simon K. Kearsley. Why do we need so many chemical similarity search methods? *Drug Discovery Today*, 7(17):903–911, sep 2002.

[48] Michael A. Skinnider, Chris A. Dejong, Brian C. Franczak, Paul D. McNicholas, and Nathan A. Magarvey. Comparative analysis of chemical similarity methods for modular natural products with a hypothetical structure enumeration algorithm. *Journal of Cheminformatics*, 9(1):46, dec 2017.

[49] Biologiske Skrifter, B V Ind, BY S Thorvald, and Rensen København. DET KONGELIGE DANSKE VIDENSKABERNES SELSKAB A METHOD OF E ST A B L ISH IN G GROUPS OF EQUAL AMPLITUDE IN PLANT SOCIOLOGY B A SE D ON SIM ILARITY OF S P E C IE S CONTENT AND ITS APPLICATION TO ANALYSES OF THE VEGETATION ON DANISH COMMONS. Technical report.

[50] Linda Sleire, Hilde Elise Førde, Inger Anne Netland, Lina Leiss, Bente Sandvei Skeie, and Per Øyvind Enger. Drug repurposing in cancer. *Pharmacological Research*, 124:74–91, oct 2017.

[51] William Thomson, Lord Kelvin, Nina Nikolova, and Joanna Jaworska. Approaches to Measure Chemical Similarity , a Review & Combinatorial Science. 22, 2003.

[52] Marco E. Turini and Raymond N. DuBois. Cyclooxygenase-2: A Therapeutic Target. *Annual Review of Medicine*, 53(1):35–57, feb 2002.

[53] David Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Modeling*, 28(1):31–36, feb 1988.

[54] Peter Willett, John M Barnard, and Geoffrey M Downs. Chemical Similarity Searching. 1998.

[55] D. H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–260, 1992.

[56] Roberto Würth, Stefano Thellung, Adriana Bajetto, Michele Mazzanti, Tullio Florio, and Federica Barbieri. Drug-repositioning opportunities for cancer therapy: novel molecular targets for known compounds. *Drug Discovery Today*, 21(1):190–199, jan 2016.

[57] Ling Xue, Jeffrey W. Godden, Florence L. Stahura, and Jürgen Bajorath. Design and evaluation of a molecular fingerprint involving the transformation of property descriptor values into a binary classification scheme. *Journal of Chemical Information and Computer Sciences*, 43(4):1151–1157, 2003.

[58] Yixiao Yun. Analysis and Classification of Object Poses : programs. *Civil Engineering*, 2005.

[59] Zhi-Hua Zhou. *Ensemble Methods Foundations and Algorithms*. ensemble_book, 2012.