Drug Repurposing using Association Rules

Sara Filipa Gonçalves Rocha

Master in Bioinformatics and Computational Biology

Department of Computer Science

2021

Supervisor:

Inês Dutra – Department of Computer Science, FCUP

Co-supervisors:

Pedro Alexandrino - Department of Chemistry and Biochemistry, FCUP

Todas as correções determinadas

pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____

Acknowledgments / Agradecimentos

Em primeiro lugar quero agradecer ao Iván por me acompanhar neste projeto desde o início, ajudou-me, apoiou-me e nunca desistiu de mim mesmo quando eu própria não acreditei. Apesar de não nos termos conhecido pessoalmente foi e é, sem dúvida, uma pessoa importantíssima e sem ele nada disto seria possível.

Em segundo lugar quero agradecer à professora Inês e ao professor Pedro por me permitirem a realização deste trabalho, pela simpatia, preocupação e apoio.

Quero agradecer à minha família pelo apoio a todos os níveis durante estes tempos difíceis que passaram. Em especial à minha irmã, aos meus pais e à Rita que foram sempre os primeiros a amparar-me.

Agradeço também às minhas amigas do Centro, que em pouco tempo se tornaram muito. Por último às minhas professoras de licenciatura que, para além de me darem emprego, sempre me motivaram e ajudaram.

A todos vocês, MUITO *OBRIGADA.*

Abstract

The main goal of this work is to present a methodology for drug repositioning using Association Rules for the discovery of new relationships between drugs and diseases.
Drug repositioning is a method for the discovery of new interactions between drugs already existing in the market or which have failed in high stages of tests for certain diseases. This process is possible since these types of molecules present polypharmacology, that is, they can have more than one target. In the last decades, drug repositioning has been a promising field due to the reduction of time spent in investigations, reduction of cost and the better efficiency compared to investigations for the creation of drugs.
In this work, 21 diseases from different groups were chosen. Afterwards, a database with 226 phenotypes and 232 drugs was created. Each one of the drugs was described by the knowledge-based method, that is, by its molecule type, mechanism of action, target and effect. The diseases were characterized using a gene prioritization methodology that creates an ordered list of genes according to their relevance to the disease.
With the application of the Association Rules, 14 rules were generated for new associations between drug and disease. From these rules we can highlight the drugs Metformin, Paclitaxel, Bevacizumab and Pegfilgrastim. Metformin is a drug commonly used in diabetes and according to the rules generated we can indicate it as having influence in mechanisms of respiratory diseases and cancer. Paclitaxel, a drug indicated for cancer, in the present results is indicated as having influence in cardiovascular and respiratory mechanisms. Bevacizumab, commonly used in cancer treatment, may have influence in cardiovascular and autoimmune (neurodegenerative) disease mechanisms. Finally, Pegfilgrastim, also indicated for cancer treatment, may have an action on immune system diseases.
This is an initial work that can be implemented in more complex databases.

Key words: Drug repurposing, Gene prioritization, Machine Learning, Association Rules, *Apriori* algorithm

Resumo

Este trabalho tem como principal objetivo apresentar uma metodologia para reposicionamento de fármacos aplicando Regras de Associação para encontrar novas associações entre fármacos e doenças.

O reposicionamento de fármacos (DR, do inglês – *Drug Repositioning*) consiste em atribuir novas indicações a medicamentos que já existam no mercado ou tenham reprovado em altos estágios quando testados para certas doenças. Este processo é possível uma vez que estes tipos de moléculas apresentam polifarmacologia, ou seja, podem ter mais do que um alvo. Esta área, nas últimas décadas tem vindo a emergir devido à redução de tempo despendido nas investigações, redução do custo e da melhor eficiência comparado com investigações para criação de fármacos.

Neste trabalho, foram escolhidas 21 doenças de diferentes grupos. Posteriormente, foi criada uma base de dados com 226 doenças e 232 fármacos. Cada um dos fármacos foi descrito por um método baseado no conhecimento, ou seja, pelo tipo de molécula, mecanismo de ação, alvo e efeito. Já as doenças foram caracterizadas usando uma metodologia de priorização dos genes que cria uma lista de genes ordenados de acordo com a sua relevância para a doença.

Com a aplicação das Regras de Associação, foram geradas 14 regras para novas associações entre fármaco e doença. Destas regras podemos destacar os medicamentos Metformin, Paclitaxel, Bevacizumab e Pegfilgrastim. O Metformin é um medicamento normalmente usado em diabetes e, segundo as regras geradas, podemos indicar como tendo influência em mecanismos de doenças respiratórias e cancro. O Paclitaxel, um medicamento indicado ao cancro, nos resultados apresentados é sugerido como tendo influência em mecanismos cardiovasculares e respiratórios. O Bevacizumab, normalmente usado no tratamento de cancro, poderá ter influência em mecanismos de doenças cardiovasculares e autoimunes (neurodegenerativas). Por fim, o Pegfilgrastim também indicado para tratamento de cancros poderá ter ação em doenças do sistema imunitário.

Este é um trabalho inicial que pode ser implementado em bases de dados mais complexas.

Palavras-Chave: Reposicionamento de Fármacos, Priorização de Genes, Aprendizagem de máquina, Regras de Associação, Algoritmo *Apriori*

Contents

List of Tables

List of Figures

Acronyms

AI	Artificial Intelligence
AR	Association Rules
CGPM	Consensus Gene Prioritization Methodology
GP	Gene Prioritization
ML	Machine Learning
NCCN	National Comprehensive Cancer Network
FDA	US Food and Drug Administration
OTP	Open Target Platform
TM	Text Mining

“A persistência é o menor caminho do êxito.”

- Charles Chaplin -

Chapter 1

Introduction

Drug repositioning (DR) also called drug repurposing, switching or reprofiling was developed in 2004 by Ashburn and Thor [10], with the main purpose of giving a new target to drugs already existing in the market. This definition also includes active substances that failed in the last stage of their approval. Overall, the process of drug repositioning consists of finding all the targets to the chemical substances [47].

DR has been present in the scientific community for a long time, being the primordial example the Aspirin [45], although it had just been defined and improved since the beginning of XXI century [47].

DR is possible due to an intrinsic characteristic of drugs called polypharmacology, which means that one single drug molecule can interact with a set of multiple targets. The presence of side effects of drugs, either being beneficial or not, were the evidence that molecules can have more than one target, or mechanism of action, as well as have other functions. This meant that drugs could even be more effective on different contexts [6].

A *de novo* drug development process can take up more than ten years, including the drug discovery process until their market availability. This, plus the increasing cost of new drugs over the last 40 years shows the importance of the investment on Drug Repositioning [66].

DR presents three main advantages compared to other drug development strategies: lower prices, less time and less risk-versus-reward [47].

There are different methods described to make drug repositioning. In 2014, Jin and Wong [45], reported six main methods for drug repositioning. The classification of DR methods is made according to two criteria: the data available about the diseases, and its final goal. DR methods are based on: phenotypic screening, target-based methods, knowledge, signatures and pathways and networks. Each method is described on the follow paragraphs as well as when it is used, etc.

Methods based on phenotypic screening use information of the disease phenotype. These are used when there is almost no information about the disease. Using this method, approximately 37% of the small molecules approved by FDA between 1999 and 2008 were discovered. The main advantage of this method is the high flexibility, allowing it to be applicable to a large number of diseases and drugs, although it is less likely to identify any mechanisms of action [45].

Target-based methods for DR are used when the chemical structure of the drug is available. These methods perform screening from drug libraries [87], and could be improved with target information.

Knowledge-based methods use information such as chemical structures of drugs and targets, drug-target networks, adverse effects, metabolic pathways and FDA approval labels [12]. The biggest advantage of these methods is the possibility to predict unknown mechanisms and improve prediction accuracy [45].

Signature-based methods use gene signatures derived from disease data, that could or not be treated. These allow the discovery of unknown drug mechanisms of action, since they allow the detection of unknown off-targets and unknown diseases mechanisms [12].

Pathway- and network-based methods involve the knowledge of omics data, signalling or metabolic pathways and protein interaction networks. The advantage of them among the others is that they reconstruct specific drug pathways to the disease which shortens the number of networks between drug and target proteins.

The most relevant factor to choose a method from one of these is the availability of data. Depending on the available databases, we can analyse similarities or affinities. For example, diseases can be described by different information such as, phenotypes, involved proteins, gene/proteins mutations, etc. Similarly, drugs can be described by the chemical structure, targets, mechanism of action, action, etc. [48].

In this work, we aim to develop a method for Drug Repositioning using information about the diseases and the drugs. For describing drugs, we will use a method of knowledge, using information such as, its targets, mechanism of action, molecule type,

etc. For describing diseases, we will use a method of gene signature getting information about which genes are involved in the disease and prioritize them.

Artificial Intelligence (AI) is a technological science that mimics and expand human intelligence, developing theories, methods and techniques to apply in many areas such as medicine and industry of cars, houses, etc. Machine Learning (ML) is a branch of AI that focuses on making predictions using classification and regression models based on known features previously learned from the training data. ML allows to overcome some barriers that AI could not overcome, having more relevance in real-word, once models could be training to explore patterns in big data in order to “learn” and does not require specific programmers to represent the knowledge [55]. In the last years many novel interactions are predicted using deep learning, i.e., more complex models such as Neural Networks models [15].

There are a bunch of ML models that could be applied in DR since the most basic ones, such as Association Rules, Support Vector Machine, Decision Trees to the most complex ones, for example Neural Networks [15]. Machine learning algorithms could be classified into two different categories: supervised (e.g. Decision Trees) and unsupervised ones (e.g. Association Rules) which takes care about the classes of the data and create groups within training data, respectively [55].

Artificial Intelligence and Machine Learning could be used at three different stages of DR: target identification, lead generation and optimization and preclinical development. To make a target identification are usually used heterogeneous datasets and are identified patterns to understand drug functions and molecular mechanisms. For lead generation and optimization stage, ML is used to optimize pipelines and models of the *de novo* drugs design. Finally, at optimization and preclinical development stage, ML/AI are used to generate predictive models to optimize drugs profiles, such as, toxicity, absorption, metabolism and distribution on the bodies [93].

This work focuses on the first application of ML/AI aforementioned, target identification finding new patterns on heterogeneous data.

1.1. Motivation

The pharmaceutical industry has a very important role once drug discovery and development can provide a rise in average life expectancy and also in life quality [74]. There are some agencies, that among other fields, they are responsible to promote public health by developing drugs. Food and Drug Administration (FDA) in the United

States, the European Medicines Agency (EMA) and the World Health Organization (WHO) within the United Nation Systems are examples of these type of agencies [4].

As drug repositioning (DR) aims to find new approaches to the drugs that are in the market or failed in highest stages of their development, it is more economic, faster and helmless. So, it became very attractive for the scientific community and for the agencies mentioned previously [6, 15, 66] and this area is increasing exponentially [47]. However, a lot of molecules/drugs are still not having application or are not well characterized with respect to all the targets or all the mechanisms of action.

Over the last decades, the number of machine learning applications on DR methods has increased exponentially and have so great interest that health agencies have launched programs of it [22]. However, obtaining effective drugs is a complex process that includes the therapeutic window where about 30% drugs fail, being the drug efficiency the main reason for drug fails in the clinical trials. For example, the success rate of drug repurposing when compared to *de novo* oncological drugs are not significant (less than 6% and 5%, respectively). Nevertheless, 30% of the new drugs approved by the FDA were indicated by drug repurposing and the time and cost are meaningfully lower [54, 69].

As mentioned above, the success of these methodologies depends on the initial data, i.e., the knowledge about diseases and drug. It is important to choose the right method and to make the best selection and treatment of the data in order to get good results [55].

The majority of studies of DR presented in the literature used data of gene expression in several cell lines after their exposed to drugs and the predictions are made using machine learning algorithms [55]. Beyond of using cell lines as diseases characterization there are other ways.

Characterize diseases by related genes and understand the mechanisms of them can be used to develop gene-specific therapies [94]. As we are witnessing the exponential growth of scientific information, tools based on text mining have been playing an important role in DR, identifying genes and relations among them and diseases [78].

Gene prioritization is a process, to rank the involved genes in specific diseases based on their contribution [36]. According to Rastegar-Mojarad *et al.* [78] the first method of gene prioritization was introduced in 2002 by Perez-Iratxeca and colleagues. Over the last years, several methods with different strategies based on the ‘guilt-by-association’ were created to make the gene prioritization. This concept means that the most relevant genes will be the ones that are most involved in biological process of interest [92, 101].

Consensus gene prioritization methodology (CGPM) is a new approach of gene prioritization that rank the genes making a mean of different tools. This strategy appears to take better results when comparing with normal prioritization tools [91].

In 2017, Tejera *et al* [91] made a study to analyse gene prioritization consensus strategy against twelve bioinformatics tools that only require the disease name for preeclampsia pathogenesis. It was also demonstrated that consensus strategy was the method with higher identification of these pathogenic genes, followed by MetaRanker method. This consensus method had the lowest average rank compared with other bioinformatics tools, which means that it detected the pathogenic genes earlier. Briefly, it was showed that consensus strategy improves the detection of the pathogenic genes.

In 2018, López-Cortés *et al.* [62], used the methodology of Tejera *et al.* [91] to validate the strategy of prioritization for breast cancer pathogenesis, against GLAD4U[48], PolySearch [25], Cipher [100], DisgeNEt [76], Génie [33], SNPs3D [106], Guildify [38], Phenolyzer [103] that are prioritization genes bioinformatics tools. The results for consensus strategy were similar, thus still being the method with the highest identification of proposed genes, however the rank of detected genes was not superior to Guildify and Phenolyzer methods. Considering both criteria recovery and ranking, consensus strategy was better than the other two, Guildify and Phenolyzer, once in initial 10% of obtained list, it recovered almost 50% more than Guildify and 20% more genes than Phenolyzer.

In 2020, Cabrera-Andrade *et al.* [18] made a study to analyse gene prioritization consensus strategy against nine bioinformatics tools – BioGraph [59], CIPHER[100], DisgeNEt [76], Génie [33], GLAD4U [48], Guildify[38], Phenolyzer [103], PolySearch [25], SNPs3D [106] - for Osteosarcoma Pathogenesis. These methods only required the disease name or the OMIM code for gene prioritization. In this study, the authors show that consensus strategy detected more pathogenic genes than other nine tools and that the number of genes and their ranking average are similar, which means the majority of the detected pathogenic genes are in top positions. They also reported that the consensus strategy has the highest percentage detecting pathogenic genes, especially in top 20% of the list with 88%, followed by Genie (80.88%), Phenolizer (72.60%) and SNPs3D (71.88%).

As said before, predictions for DR are mainly made by using machine learning algorithms that can be supervised or unsupervised algorithms. However, to start the process of drug repositioning, is more frequent use unsupervised methods, for example Association Rules (AR) that are easier to understand [17, 35] and obtain high accuracy [14].

1.2. Objectives

This research proposes a novel method for drug repositioning application, using AR to find interactions between drugs and diseases which are described by two methods based in Knowledge and Gene-Signature, respectively. Six main objectives are proposed:

1. Select, collect and process data about Drugs and Diseases;
2. Describe each disease from the selected data using a consensus gene prioritization method and support results with literature;
3. Describe each drug with the data and literature and associate them to diseases;
4. Explore the data and use Association Rules (*Apriori* algorithm) for the discovery of new drug-disease interactions;
5. Analyse the new discoveries;
6. Propose improvements to the methodology.

1.3. Our Approach

The first goal to collect the data was defining some criteria for the selection of diseases. All diseases should be polygenic, have a high incidence on population. Furthermore, they should be grouped by type (e.g., autoimmune, neurological, cancers, etc.) and have some correspondences between them, particularly with cancers (e.g., Psoriasis and Skin Cancer).

From the initial diseases, we should expand the disease set, using online free databases, such as *Drug Bank, ChEMBL, Experimental Factor Ontology.* There are many methods for the prediction of drug-disease interactions using *omics* and clinical data. We described diseases using a Signature-Based method and drugs using a Knowledge-Based method.

We collected data about drugs from different sources in order to obtain fairly complete information about targets, stage of development, and indicated diseases. Diseases were described by a list of prioritized genes using a method of consensus strategy described by Tejera [91] that, as seen, on previous point, was proved to be the most efficient in the recognition of gene-disease association.

After these steps, we should do a description of drugs and diseases. The drugs should be described according to the data and the literature. For diseases, the most

relevant genes for each group type should be revealed and characterized in terms of mechanisms of action. Furthermore, each disease should be compared with others of same type and with cancers, to identify similarities.

The most important point of this work is the application of a Machine Learning model. We should apply Association Rules, more precisely with the *Apriori* algorithm [17, 35]. The data should present each disease associated with all the drugs indicated for its treatment and the results should be rules containing drugs that could be indicated for the same disease.

Combining the information about diseases, drugs and the association rules we should make assumptions and reveal new discoveries.

1.4. Contribution

Drug Repositioning is an important topic nowadays, and it is almost indispensable with the amount of information that have been collected recently. However, getting efficient medicines is a hard process, that requires time, computational power and a numerous studies and reliable information. The success rate still low and there are a lot of improvements to do.

The main objective of this work is to find new associations between drugs and diseases applying Associations Rules. To achieve this, we must overcome two problems: (1) the description of drugs and diseases using two different methods, namely, Knowledge-based and Signature-based, and, (2) identify similarities between diseases based on their prioritized gene lists.

The contribution of this work is meaningful once we should identify drugs-diseases associations that can be tested in future work.

We present supplementing information that could be used in future drug repositioning projects. The description of top genes in diseases types and comparison of each disease leads to a better understanding of the most relevant mechanisms involved on each group. Furthermore, it helps to evaluate if the consensus prioritization strategy had good performance.

1.5. Outline

This dissertation contains 5 chapters to present all the work performed over the last few months.

Chapter 1 presents an introduction about what is drug repositioning and the importance of this field. Are also explained the objectives and our approaches, explaining the reasons why each step was applied, as well as the contribution of this work to the scientific community.

Chapter 2 presents the main concepts that are used during this work and summarizes the main related work.

Chapter 3 presents the materials and methods in this work, performed to achieve the principal objective, find new association between drugs and diseases. It starts with a small diagram which shows succinctly the main steps of our methodology, and then explain step by step which one of them. First, why and how the data were selected, its preparation and lastly, the methodology of Association Rules (the *Apriori* algorithm).

In chapter 4, a brief summary of the data and results is presented. Genes and drugs are described, it is showed the comparison of diseases according to the CGPM and the rules obtained with the Association Rules method. Also, all of the results discussed during the chapter.

Finally, chapter 5 summarizes the accomplishments of the work, showing the positive aspects and the limitations of this research and also presents a few ideas for future work.

Chapter 2

Background

As said in the previous chapter, this work aims to find new possible associations between drugs and the selected diseases. To achieve this, our approach is to implement Association Rules on a dataset with the association between drugs and diseases. Diseases were primarily described by a prioritized list of genes in order to compare them. These lists were generated with a consensus gene prioritized methodology.

Before going any further, it is important to define some basic concepts that will be used during this dissertation.

2.1. Basic concepts

2.1.1. Gene prioritization

Gene prioritization is a process that aims to discover the diseases-related genes ranking them according to their likelihood of being associated with the diseases [36]. Due to the large and scattered information, this methodology uses computational methods that focus on text mining (TM) strategies [36]. TM is defined as the automatic information discovery from unstructured data. This comprises three main tasks: selection of important documents, extraction of relevant information and find important relations between information [8].

Over the last years, several methods with different strategies based on the 'guilt-by-association' principle were developed which means the most relevant genes will be the ones that are most likely involved in biological process of interest [92, 101].

Additionally, we can combine several tools for GP using a consensus strategy, since there is evidence for supporting that a consensus has better results than any of their components individually [92, 101]. Consensus gene prioritization is a computational gene prioritization strategy that, as the name implies, ranks genes creating a consensus between several gene prioritization tools [18, 30, 91].

2.1.2. Association Rules

Association rules (AR), also known as Association Rule Mining (ARM), is a well-known data mining technique which is widely used for the discovery of interesting patterns in large datasets. This technique requires several parameters, e.g., support, confidence, lift and length minimum, among others, and through them, generate rules about data associations. AR are efficient algorithms that can be applied in several fields, such as, market, medicine, in biology, streaming services, etc. [14].

AR is commonly used in first stages of drug repositioning given that it is an unsupervised algorithm, with results having a certain degree of explain ability [17, 35].

2.2. Related Work

Over the last years, methodologies of DR have been widely applied to many therapeutic areas. One of the most challenging and dominant tasks is predict the drug targets, highly related to diseases in order to the drug be effective [93]. Below are summaries of some studies (2.2.1) of different drug repositioning approaches and (2.2.2) some ways of describing diseases by genes mostly of gene prioritization methods. Table 1, on page 16, at the end of this chapter, are briefly presented the studies presented on point 2.2.1.

2.2.1. Different Approaches of DR

Kuo *et al.* in 2009 [56] implemented the *Apriori* algorithm for detection of ADR (Adverse Drug Reaction) with drugs combination. They collected information about patient's characteristics, which treatments they were subjected to, their primary diagnosis and the ADRs. The results and databases used were not available.

Bresso *et al.* [15] in 2013, used a Knowledge-based method to predict side effects of drugs. They extracted data from SIDER, DrugBank and PDB. To achieved the goal, they used two different methodologies Inductive Logic Programming (ILP) and Decision Trees (DT). From DrugBank and PDB they extracted information about drugs such as targets, protein-protein interactions and pathways. From SIDER they extracted side effects of drugs and they made a hierarchical cluster. Next, they associated drugs and their side effects information and implemented ILP and DT. They trained models to predict side effects profiles and they found that ILP models have a higher sensitivity than DT models, i.e., ILP predicts side-effects more often than decision trees. ILP also show that background knowledge is well exploited during rule induction so it can be used in drug repositioning in order to anticipate side effects.

Jung and Lee in 2013 [49] used two different methodologies to descried data, Knowledge-based and Signature-based and extracted it from National Health and Nutrition Examination Survey (NHANES) dataset. They applied AR with a FP-growth algorithm. First, they selected a set of 26 different diseases and according to the acquired information/data they made a similarity between diseases. They found that the gout and heart diseases pairs were positioned at high ranks in this, but hardly shared drugs for the disease pairs yet and proved the results with the literature.

In 2014, Boutorth *et al,* [14] developed a new method of Association Rules which they named of Grammatical Evolution Association Rules Method (GEARM). They implemented knowledge and structure-based methods of DR and pulled information from DrugBank, OMIM and CTD. Their aim was to find new pairs (Drug, Disease) and they obtained a data of 288 drugs and 267 diseases. They found three new pairs: (i) (Carvidol, Diabetes Miellitus); (ii) (Dihomo-linolenic acid, Breast cancer) and (iii) (Dihomo-linolenic acid, Colorectal Cancer).

Min Oh *et al.,* in 2014, [70] used two different methodologies of DR Pathway- and Network-based and Knowledge-based. They pull information from several databases: Online Predicted Human Interaction Database, Pathway Interaction Database, DrugBank, OMIM and Comparative Toxicogenomics Database (CTD). They implemented a network model and compared with Random Forest (RT). Their model obtained high sensitivity and specificity with an AUC of 0.855 and better results than Random Forest. They proposed two new interactions between drug diseases: Propanol with Cancer and Telmisartan in Alzheimer.

Asdame *et al*. [7], in 2020, made a research to find drugs that inactivate B-cells, whose dysregulation could lead to Autoimmune diseases. They screened 500 kinases using a database named BidingDB and they identified 22 kinases with an interesting pharmacological profile in B Cell activation/inhibition. Then, they used a Signature-based

method and resorted to Protein Data Bank (PDB) to extract kinases structures, as a binary fingerprints, and reduced their kinases to 10 due to not having all templates on the database. They compared their fingerprints kinases with fingerprints of the full PDB, calculating similarity sing a PubChem Score Matrix Service. Manually, they used PubMed and Open Target Database to link kinases targets with diseases and then PubChem to link drugs with diseases. They implemented a drug-target-disease model. Finally they found network that *Ibrutinib* drug which is a VEGFR2 inhibitor with a high therapeutical index in B-cells, acting beyond on Bruton's tyrosine kinase BTK [7].

Bo-Wei Zhao *et al.* in 2021, [108] proposed a new approach to discover new interactions between drugs and diseases creating a large-scale graph (LSG). They used three methods of DR on their work: Knowledge-based, Signature-base and Pathway- and Network-based. For that, they obtained data from DrugBank and STRING database. They implemented several models such as Large-scale graph, Logic Regression (LR), K-nearest neighbour (KNN), Decision Trees (DT) and Random Forest (RF). From that work, they obtained several results: (i) LR were not well succeeded due to the high number and complexity of input features; (ii) KNN were also not well succeeded due to a fusion of some attributes in earlier stage; (iii) DT and RF obtained good results of accuracy, especially RF; (iv) LSG has the best performance compared to other models; (iv) Using LSG they found that Clozapine could be used to treat several phenotypes of schizophrenia.

2.2.2. Genes related to Diseases

Discover the disease-caused genes is an important point on the DR research so that several studies implemented ML classifiers for example, Decision Trees (DT) to achieve it [93]. As explained in previous chapter, gene prioritization has been more widely used. Above there are described some methods of this methodology.

In 2011, Chen *et. al.* [24] proposed a new method of an *in silico* prioritization based on online databases. They reported four gene associated to Parkinson Disease: UBB, SEPT5, GPR37 and TH.

In 2016, Cruz-Monteagudo *et al.* [30] studied a consensus strategy of prioritization genes also for Parkinson's disease and then compared to an experimental microarray data. The consensus made a combination of three different approaches: *Limma*, Machine-learning and Weight Gene Co-expression. From the 102 samples of Parkinson diseases and healthy control they provided a set of 50 genes where 6 of the top 10 genes were directly associated with Parkinson disease. Eight genes (CCNH, DLK1, PCDH8, SLIT1, DLD, PBX1, INSM1, and BMI1) related to the biological process

in disease were found, i.e., potential biomarkers or therapeutical targets were also discovered.

In Tejera *et al* [91] study about gene consensus prioritization in preeclampsia disease, they obtained a set of 476 genes, with a lot of communities connected with VEGF- signalling pathway, being the VEGF, KDR, FLT1 the most relevant to the diseases. They also reported three more genes, PAK2, CD247 and HSP90 to experimental approaches.

The study of López-Cortés *et al.* [62], used the methodology of Tejera *et al.*[91] to validate the prioritization strategy, using breast cancer pathogenesis data. The results showed a high connection of disease to several genes like ERBB, mTOR, VEG, MAPK, PI3K-AKT and HIF-1 but the genes with the highest ranking were TP53, ESR1, BRCA2, BRCA1 and ERBB2.

In 2020, Cabrera-Andrade *et al.* [18] created a list of prioritized genes for osteosarcoma. This list was composed by 553 genes lead by TP53 gene. The genes involved in activities such as metastasis (MMP2 and MMP9), and in DNA repair pathways (ATM, RAD51, ATR and CHEK1) were also presented in the list with a good rank.

2.2.3. Comparison of studies with our work

For this work we proposed to use a methodology based on knowledge and signature to describe drugs and diseases, respectively. Observing the studies presented in this chapter, these are the most used methodologies, as there is more available data and in general the results are good.

Asdame *et al*. [7], in 2020, used several databases, one of them was OTP. In this work we used this database to describe and associate drugs to diseases. To describe diseases, we proposed using a gene prioritization methodology, described by Tejera that obtained a good performance by of López-Cortés *et al.* [62].

The model that we used for this work were Association Rules with *Apriori* algorithm as used by Kuo *et al.* in 2009 [56]. As Jun and Lee [49] we proposed selected a set of different diseases as a start point of this work. Also, instead of calculate the similarity between diseases we compared, manually, diseases with the lists of prioritized genes obtained previously. Our obtained rules were different from Boutorth *et al,* [14] because they aim not to pairs (disease, drug), but (drug, drug) based on diseases association similarity.

Table 1. Revision of drug repositioning work using machine learning models.

YEAR, AUTHOR	USED DATABASES	METHODOLOGY	ML MODEL	RESULTS	REFERENCE
2014, Min Oh *et al.*	Online Predicted Human Interaction Database	Pathway- and Network-based and Knowledge-based	Network	They obtained a high sensitivity and specificity with their integrative genetic network and better results compared with the implementation of RF. Proposed two new interactions: (i) Propanol could be used in cancer (ii)Telmisartan used in Alzheimer	[70]
	Pathway Interaction Database				
	DrugBank				
	OMIM				
	Comparative Toxicogenomics Database (CTD)				
2013, Jung and Doheon	National Health and Nutrition Examination Survey (NHANES) dataset	Knowledge- based and Structure-based	Association Rules - FP-growth algorithm	Found that the gout and heart diseases pairs were positioned at high ranks in this, but drugs hardly shared for the disease pairs yet	[49]
2009, Kuo *et al.*	Not available	Knowledge - based	Association Rules – Apriori Algorithm	Not available	[56]
	Drug Bank				

2014, Boutorh *et al.*	Online Mendelian Inheritance in Man (OMIM) Comparative Toxicogenomics Database (CTD)	Knowledge-based and Structure-based	Grammatical Evolution Association Rules Mining (GEARM)	(i)Carvidol already related to Inflammatory bowel, Rheumatoid Arthritis, found a new pair, Diabetes Mellitus (ii) Dihomo-linolenic acid associated to Breast cancer (iii) Dihomo-linolenic acid associated to Colorectal Cancer	[14]
2020, Asdame *et al.*	BidingDB Protein Data Bank (PDB) Open Targets Platform PubChem	Structure-based	Network	*Ibrutinib* drug is a micro molar VEGFR2 inhibitor with a high therapeutical index in B-cells that acts beyond on Bruton's tyrosine kinase BTK	[7]
2013, Bresso *et al.*	SIDER DrugBank Protein Data Bank (PDB)	Knowledge-based	Decision Trees (DT) Inductive Logic Programming (ILP)	ILP models have a higher sensitivity than DT models, i.e., ILP predicts side-effects more often than decision trees. ILP also show that background knowledge is well exploited during rule induction.	[15]
2021, Zhao *et al.*	DrugBank STRING database	Knowledge-based; Structure database; Pathway- and Network-based	Large-scale graph (LSG) Logic regression (LR) K-nearest neighbour classifier (KNN) Decision Trees (DT) Random Forest (RF)	(i)LR were not well succeeded due to the high number and complexity of input features (ii)KNN were also not well succeeded due to a fusion of some attributes in an earlier stage (iii)DT and RF obtained good results of accuracy, especially RF (iv)LSG has the best performance compared to other models (iv)Clozapine could be used to treat several phenotypes of schizophrenia, has hypnotic effect	[108]

Chapter 3

Material and Methods

In order to achieve the main objective proposed, several steps were made and they explained during this chapter. These steps are explained one by one and presented chronologically.

Before even to start, figure 1 resumes all the used processes in a schematic way to a better understanding of this work.

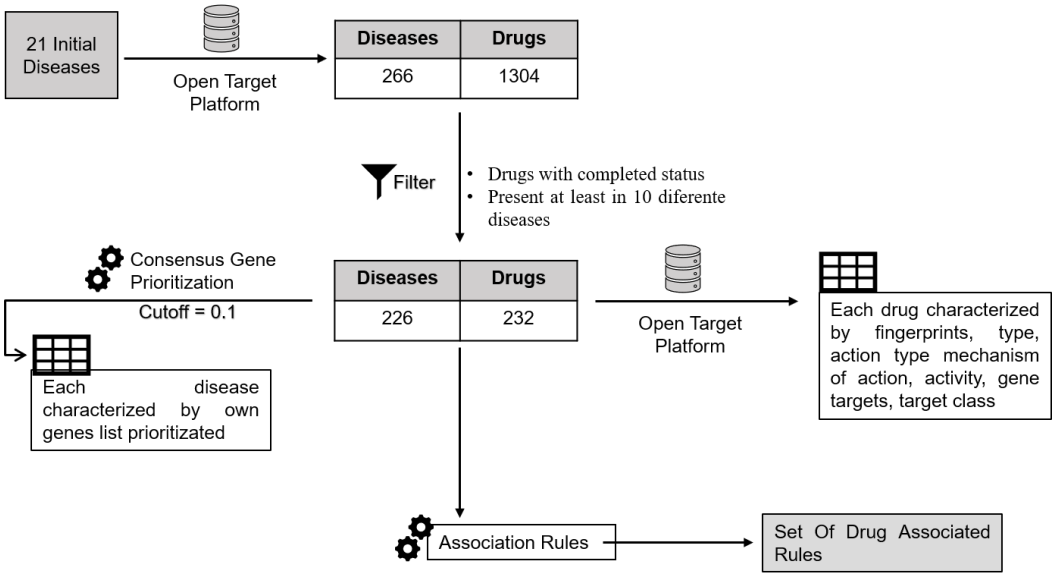

Figure 1. Drug repositioning methodology used on this work. The diagram resumes all the methodologies since the first twenty-one diseases, passing by all steps to obtained a data with 226 diseases and 232 drugs. Then, how drugs and were characterized and final the application of the Association Rules.

3.1. Data collection and preparation

In section 1.3, we stated that a set of diseases satisfying four criteria were selected: polygenic diseases, with high incidence in the population, grouped by type (e.g., autoimmune neurological, cancers, etc.) and with some correspondences between them, particularly with cancers (e.g., Psoriasis and Skin Cancer). After this, more data were found on free online databases. We utilized a compendium database, named Open Target Platform (OTP), that gives information from several databases such as, ChEMBL, Experimental Factor Ontology, among others.

OTP provides information about drugs such as: status (if it is being studied, removed from studies or completed the phase of the study), type (small molecule, protein, antibody), mechanism of action, action type (inhibitor, agonist, an opener, a blocker) (Annex II), activity (negative/positive modulator or other), gene target and target class (antibody, small molecule, protein) (Annex III).

Besides that, OTP, gives different phenotypes of the same diseases and for each one, gives all information presented before. So, it is important to note that diseases that were obtained could correspond to different diseases (e.g., *Cardiac Arrest* and *Arterial Fibrillation and Alzheimer*) but also to different phenotypes of a specific disease (e.g., *Beta Thalassemia Major* and *Beta Thalassemia*). However, each one of them was treated as different and unique disease. On other hand, drugs presented all status of clinical trials even the ones whose study was not finished.

3.1.1 Data characterization

As explained above, in previous chapter, drugs and diseases were characterized using different methods, Knowledge and Gene-Signature-based methods, respectively. To describe drugs, were used the information provided by OTP: status, type, action type, activity, targets and target's type.

On the other hand, diseases were characterized using a CGPM tool described by Tejera *et al.* [91], as described next.

3.2.1. Consensus Gene Prioritization

The software that was applied combines the scores obtained by several independent methods – BioCart, Candid, GLAUG4, PolySearch, CIPHRE, Guildify,

DISGENET, Geneprospector, GENIE, SPNS3D, GeneDistiller, and MetaRanker, using a methodology for consensus gene prioritization, as the following equations explain.

Equation (1) calculates the geometrical mean between the average score of genes predicted by each independent method and the normalized scored based in all combined methods. In this equation, *i* represents each ranked gene and the gene given by each method represented *j*. $GeneN_{i,j}$ represents the normalized score of gene *i* in method *j*. Also, n_i represents the number of methods which predicted gene *i*.

$$Genei = \sqrt{\left(\frac{ni-1}{12-1}\right)\left(\frac{1}{j}\sum_j GeneNi,j\right)}$$

(1)

All genes are sorted according to $Gene_i$ values, and according to the average calculated with equation (1). This sorting produces a ranking that is normalized. After this normalization the $ConsenScore_i$ for each gene is obtained with the equation (2).

$$CosenScorei = \left[\left(\frac{ni-1}{12-1}\right) + \left(\frac{1}{j}\sum_j GeneNi,j\right)\right]/2$$

(2)

If gene *i* is only predicted by one method, the result of this formula is zero. If two genes have the same normalized score, they have the same ConsenScore. Since the final normalized list used to be very long, a complementary method was implemented to reduce the noise and keeping the most relevant genes for that specific disease.

$$Ii = \frac{TPi}{FPi+1}\ ConsenScorei$$

(3)

Equation (3) shows an index created to overcome this problem and makes a cut-off. The maximum value of this formula shows the maximal compromise between True Positives (TP) and false positive (FP) rate that are really important to keep with $ConsenScore_i$.

At the end of this step, each disease was described by a list of ranked genes. These lists were used in order to find which genes are more related to each diseases type according to this method.

3.2. Data Exploration

After the data characterization was complete, we made an exploration of it.

First, we explored our data about drugs, observing which ones are most used, which type are them and in which diseases are they used for.

From the diseases of same type, we observed the most important genes according to the prioritized lists. Additionally, we used their top 20 prioritized gene lists and we compared them in pairs. From these comparisons we found how many and which genes are share in top 20 of same type diseases.

3.3. Association Rules - Machine Learning Application

The third step was the application of AR which description of how it works and the meaning of each parameter are exposed after on 3.4.1. In this step, we also made a small description of the drugs presented in the final rules obtained.

3.3.1. Association Rules

Association Rules are a machine learning technique that finds patterns within a large dataset and from them creates rules. Each rule can have a unique item or a set of items, a specific characteristic from this model [3]. A few metrics are used to understand the strength of the rule:

- *Support* or C*overage* – refers to how frequent an itemset is in all the rules. With is value is possible to identify which rules we should consider for analysis. Support is calculated as equation (4) being X and Y two different rules' item set.

$$Support(X \rightarrow Y)x = \frac{Rules\ containing\ both\ X\ and\ Y}{\text{Total number of Rules}} \quad (4)$$

- *Confidence or Accuracy* – defines the likeliness of occurrence of consequent on the rule given that the rule already has the antecedents. Confidence is calculated as equation (5) being X and Y two different rules' item set.

$$Confidence(X \rightarrow Y)x = \frac{Rules\ containing\ both\ X\ and\ Y}{\text{Rules containing X}} \quad (5)$$

- *Lift* – calculates the conditional probability of occurrence of Y given X, being X and Y two different rules' item set. Lift is calculated by following equation (6).

$$Lift(X \rightarrow Y)x = \frac{\frac{(Rules\ containing\ both\ X\ and\ Y)}{(Rules\ containing\ X)}}{\text{Fraction of Rules containing Y}}$$

(6)

3.3.1.1. *Apriori* Algorithm

The *Apriori* algorithm of association rule mining is a base method, that boosted data mining research. It requires multiple passes over the data and considers any subset of a frequent item set to also be a frequent item set [1]. Applying *Apriori* algorithm requires some parameters:

- **Support minimum** - select the items with support values greater than the value specified by the parameter
- **Confidence minimum** - filters those rules that have confidence greater than the threshold specified by the parameter
- **Lift minimum** - specifies the minimum lift value for the short-listed rules
- **Length minimum** - specifies the minimum number of items in the rules

We obtained a set of rules for each combination of the parameters.

To apply this machine learning method, diseases and drugs were associated as shown in the table below (table 2).

Table 2. Transaction data's model for Association Rules application

TRANSACTIONS	DISEASES	DRUG 1	DRUG 2	DRUG 3	DRUG 4	…
1	Disease 1	Drug A	Drug B	Drug C	Drug D	…
2	Disease 2	Drug E	Drug A	Drug F	**NONE**	…
3	Disease 3	Drug A	Drug B	Drug G	Drug H	…
…	…	…	…	…	…	…

Each row of table 2 represents the set of drugs associated with each disease. Using this format of data the model could find associations between different drugs, generating sets of rules. This is key to achieve our main objective.

To apply the *Apriori* Algorithm, four parameters had to be defined, as it was explained in section 3.3.1.1. Different parameters were combined on 9 different tests, as represented in table 3.

We tested three different supports: 0.1, 0.2 and 0.08. They were calculates dividing the frequency with which a drug appears in the transaction by the total number of transactions. So, when it is testing a support of 0.1 (25÷226 ≈ 0.1), 0.2 (45÷226 ≈ 0.2), 0.08 (18÷226 ≈ 0.08), it means that a drug appears in the transactions/diseases 25, 45, 18 times, respectively.

Table 3. Nine tests with different parameters on Apriori algorithm

PARAMETERS	MINIMUM SUPPORT	MINIMUM CONFIDENCE	LIFT	MINIMUM LENGTH
Test 1	0.2	0.8	2	2
Test 2	0.1	0.8	2	2
Test 3	0.2	0.6	2	2
Test 4	0.1	0.6	2	2
Test 5	0.08	0.8	2	2
Test 6	0.1	0.6	3	2
Test 7	0.1	0.8	3	2
Test 8	0.1	0.8	2	3
Test 9	0.1	0.6	2	3

Also, all of these testes were applied into three different datasets (i.e., three versions of the same dataset). The first version (V1) is the *original* data, previously described with 226 diseases and 232 drugs. The second version (V2) presents all complete status drugs that are presented in all diseases containing 266 diseases and 1304 drugs. Finally, the third version (V3) covers all the drugs in all development stages and all diseases, representing 252 diseases and 493 drugs.

After obtaining the set of rules, we made a description of drugs that are present on the strongest ones based on the drugs data collected and on the literature.

3.4. New Drugs-Disease interactions

This step was made after the analysis of the results obtained on 3.2 and 3.3.

In the first step we had a list of 21 diseases related to own type disease (neurologic, cancer, autoimmune, etc.). The diseases achieved in a later stage, were also associated to respective disease type.

All of diseases obtained were characterized using a method of consensus gene prioritization.

After this step, we compared obtained gene lists of diseases from same type to observe which ones share more genes and which genes are most common/relevant on them.

Association rules are applied in a dataset with diseases-drugs associations. As we have a large spectrum of diseases, we could predict new associations between drugs and diseases. As explained in the previous chapters, the particularity of this work is not only focus in a particular disease but in a heterogeneous group of diseases where each one has different phenotypes (although every single phenotype is treated as a specific and unique disease).

A single rule shows that a Drug A can be used to the same diseases as Drug B. So that, after making a combination of AR with the information obtained about genes and drugs, such as characterization and literature description, it was possible to indicate new drugs-disease interactions. Below, (A1) represents how a rule is showed after the AR algorithm application.

(A1) Drug A → Drug B

3.5. Summary

To summarise, this chapter presents chronologically all steps of this work since the data collection to the results interpretation. Moreover, here it is explained how to gene prioritization software works, as well the *Apriori* algorithm and how Association Rules were obtained.

Chapter 4

Results and Discussion

Throughout this chapter we will present the results obtained in each step described on Chapter 3, along with a brief discussion. Results will be presented in the same order as the steps on previous chapter.

4.1. Data collection and preparation

From the four criteria defined above, twenty-one diseases were chosen and are presented in table 4, associated with their group type.

Table 4. Set of twenty-one initial selected diseases associated to the corresponding group type.

DISEASE	GROUP
Alzheimer	Neurological
Schizophrenia	Neurological
Parkinson	Neurological
Epilepsy	Neurological
Rheumatoid Arthritis	Autoimmune
Psoriasis	Autoimmune
Multiple Sclerosis	Autoimmune
Lupus Erythematosus	Autoimmune
Pulmonary hypertension	Pulmonary
Asthma	Pulmonary
Cardiac Arrythmia	Cardiovascular
Obesity	Cardiovascular
Diabetes Mellitus	Cardiovascular
Anemia	Cardiovascular
Uterine cancer	Cancer
Prostate cancer	Cancer
Skin cancer	Cancer
Lung cancer	Cancer
Intestinal cancer	Cancer

Breast Cancer	Cancer
Brain Cancer	Cancer

According to World Health Organization, the most common cancers in Portugal are breast, prostate and lung cancers. Also, according to Portuguese Society against cancer, prostate and uterine cancers have the biggest incidence in Portuguese men and women, respectively [5].

The association between the brain and the intestine have been increasingly studied and proved, therefore were also selected intestinal and brain cancer [19]. The heterogeneous profile among the autoimmune diseases (Rheumatoid Arthritis, Psoriasis, Multiple Sclerosis and Lupus Erythematosus) was the reason of their selection [27].

From OTP we collected more phenotypes of previously selected diseases. Each phenotype was treated as an individual and unique disease. So, in this step we obtained a database with 266 diseases and 1304 drugs. Furthermore, we also collected information about drugs about their type, their action, activity, their target and target type (Annex II and III).

Afterwards, were processed the data, using two criteria based on drug data: only drugs with completed clinical trials status, and only drugs which appear at least in ten different diseases were selected. At the end, we obtained a dataset with 226 diseases (Annex I) and 232 drugs.

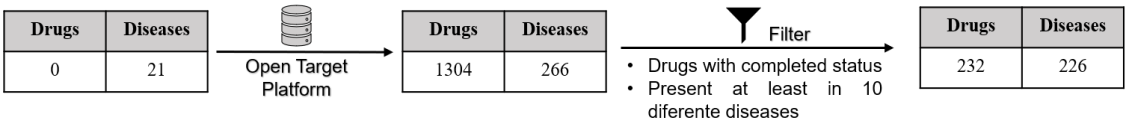

Figure 2. Data Collection and preparation.

4.2. Data exploration

As said in the previous chapter, after the data collection and preparation, we made an exploration of what were obtained. So, in this point are present the data characteristics related to drugs and diseases after their characterization.

4.2.1. Drug data exploration

In the collected data about drugs, small molecules are the predominant drug types used in drugs development (see Figure 3). Small molecules are compounds with low molecular weight which have the capacity of modulating biochemical processes to prevent, treat or even diagnose diseases. There are a few reasons for these drugs being so attractive to the pharmaceutical industry. One reason is the low weight and the simple structure making their pharmacodynamics and pharmacokinetics more predictable. Also, the protocols of their development are also simpler than other molecules. Another reason is their stability, allowing the oral intake making it more attractive for the patient. This allows these molecules to be more accessible on the market [68].

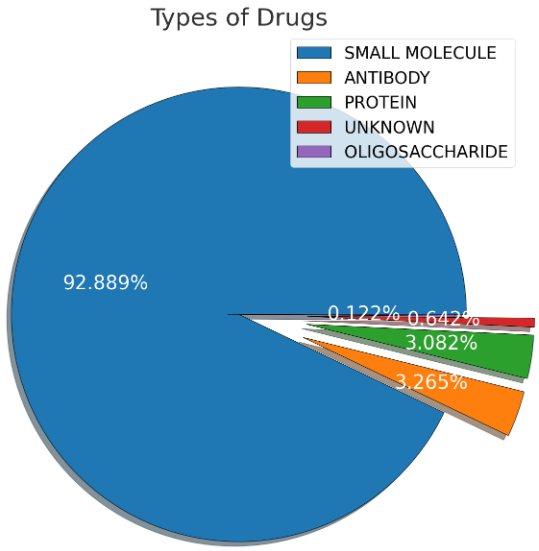

Figure 3. Types of drugs on data. Pie chart showing the proportions of developed drugs presented in the data.

In Figure 4, we see the top 6 of the most utilized drugs in our database. Starting by the most used drug, we have in the first position Paclitaxel (68 diseases), followed by Bevacizumab (58 diseases), then Pembrolizumab (57 diseases), Metformin (48 diseases), Dexamethasone and Irinotecan (both with 40 diseases).

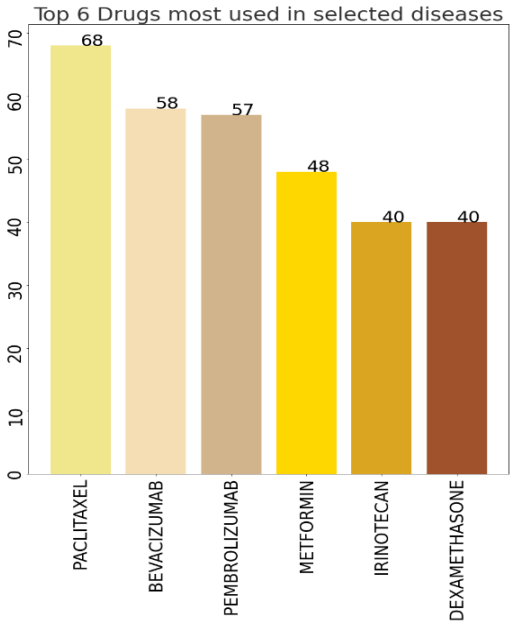

Figure 4. Most used drugs on the data. The bar graph shows the top 6 of most applied drugs in the selected set of diseases.

All top 6 drugs are indicated to cancers, mainly inhibitors and with exception of Metformin all of them only have one specific target. Paclitaxel, Metformin, Dexamethasone and Irinotecan are small molecules. Apart from Dexamethasone all of them are negative modulators with a specific but different target. Bevacizumab and Pembrolizumab are both negative modulators antibodies.

4.2.2. Disease group characterization

We had sets of 40, 12, 19, 38 and 117 diseases that corresponds to cardiovascular, pulmonary, autoimmune, neurologic and cancer types, respectively.

Each disease was described with a list with ranked genes sorted according to its influence on disease.

In table 5 each disease type is related with the three most frequent genes (“Top 3 genes complete list”) associated to respective frequency, i.e., number of diseases where each gene is present and related with the three most frequent genes on top 20 ranked genes list (“Top 3 genes on top 20”) and respective frequency. Following, a small description of each disease type and a brief presentation of genes are presented on tables below.

Table 5. Diseases type with respective top 3 of genes list complete and of top 20 genes.

DISEASES TYPE	TOP 3 GENES (COMPLETE GENE LISTS)	FREQ	TOP 3 GENES (TOP 20)	FREQ
CANCERS (117 DISEASES)	TP53	86	TP53	52
	EGFR	76	EGFR	33
	PIK3CA	73	CDKN2A	29
AUTOIMMUNE (19 DISEASES)	TNF	16	TNF	16
	HLA – DRB1	15	HLA – DRB1	15
	IL10 / IL12B/ CTLA4/ IL2RA	14	IL10 / IL12B/ CTLA4/ IL2RA	14
NEUROLOGIC (38 DISEASES)	KCNQ2	11	KCNQ2	6
	GABRB3	10	SCN1A	5
	MECP2/ CDKL5	10	CDKL5	5
CARDIO VASCULAR (40 DISEASES)	TNF	19	PPARG	6
	IL6	18	ADIPOQ	6
	ACE	16	SCN5A	6
PULMONARY (12 DISEASES)	ADRB2	5	IL5	3
	TGFB1	5	IL4	3
	NOS3	5	IL13	3

4.2.2.1 Cancers

Cancers are related with sets of cells with genetic alterations growing and multiplying mostly due to errors on growth-regulatory genes. There are two classes of these growth-regulatory genes, proto-oncogenes or tumour suppressor genes, that differ from each other by the way the control the cell growth. As the name implies, tumour-suppressor genes control the cell growth, repairing DNA errors or signalizing the cell for apoptosis (death). On the other hand, proto-oncogenes help the cell growth and when mutated they are called oncogenes [57, 94].

In table 5, we observe that P53 and EGFR are the leader genes of prioritized lists, along with PIK3CA and CDKN2A genes.

P53 is one of the most important tumour-suppressor genes whose transcribed protein is involved in a mechanism inducing cells to apoptosis (cell death). Mutations on this gene are the most related to cancer predisposition [80].

EGFR (epidermal growth factor receptor) is a proto-oncogene that when mutated (oncogene) leads to the overexpression of the related protein also called EGFR. This overexpression is related to an uncontrolled cell growth [75].

PIK3CA and CDKN2A are kinases coding genes of Phosphatidylinositol 3-kinase and cyclin-dependent kinase inhibitor 2A, respectively. Kinases are a large family of enzymes that catalyses a protein phosphorylation, a critical mechanism to regulate several cellular functions such as cell cycle, growth, differentiation, apoptosis, etc. There are numerous types of kinases and a single deregulation on their function can lead to no

phosphorylation or an aberrant phosphorylation, which could lead to different cancers [28]. PIK3CA and CDKN2A are considered oncogenes that has been reported with several mutations in a variety of cancer types. As their proteins have an important role in tumour suppressor mechanisms, mutations in these genes lead to an uncontrolled cell growth.

In summary, these genes are associated to the cell cycles control which supports the essence of cancer, a deregulation on a multiplication, growth of cells and a failure signing cells to apoptosis.

4.2.2.2. Autoimmune diseases

Autoimmune diseases are diseases which the immune system attack the own body. These diseases are known as a multifactorial pathogenesis, i.e., there are genetic and environmental factors involved in their expression. These diseases share several risk genes suggesting a common pathways is involved which could be a good outset for therapies [20].

Some studies say that approximately half of the autoimmune diseases genetic susceptibility are related with HLA (Human Leukocyte Antigen) region and this strong association have been reported for decades [20, 85]. As shown in table 5, a gene of these region is reported as the second most important.

HLA region is the densest and more polymorphic region in the human genome. Among many genes, this region encodes molecules with important roles in the immune system. This region can be divided into five classes (extended class I and II and classics, class I, class II and class III), especially, with components of class II (HLA-DRB1, HLA-DQA1, HLA-DQB1). The components of HLA activate T-cells (also called T helper cells) which play an important role in the recognition of threats that immune system should respond to [85].

Other genes, such as, PTPN22, STAT4, TNF, IL-1, IL-6, IL-4, IL-5, CTLA4 and PRF are also related to autoimmune diseases, in particular, to Rheumatoid Arthritis and Systemic Lupus Erythematosus [20].

TNF (tumour necrosis factor) and its protein, cytokine TNF, has an important role in signalizing cells to induce an inflammatory response. Therapies that focus on this gene using antibodies have shown a good performance [34].

IL (interleukins) proteins are a family of 18 molecules that communicate to immune cells to induce them to divide, differentiate and, similarly to TNF protein, they have an essential role in the inflammatory system [13].

CTLA4 (cytotoxic T-lymphocyte antigen-4) is a protein coded by CTLA4 gene that regulates the immune system, transmitting an inhibitory signal to T-cells and stopping the inflammatory response [81].

All of the reported genes on the literature are shown on top 3 of autoimmune diseases, which means a good performance of the prioritization consensus strategy.

4.2.2.3. Neurologic diseases

Neurologic diseases are specially challenging for scientific community to find risk factors for onset and progression, due to their very heterogeneous profile in terms of frequency, age, progression and aetiology [60]. According to consensus gene prioritization strategy, the most important genes are KCNQ2, GABRIB3, MECP2, CDKL5 and SCN1A.

KCNQ2 gene belongs to a family of genes that deliver instructions for making potassium channels. These channels play important roles in the determination of the resting/action potentials in neural cells [97]. When a deregulated potassium currents exists in neurons, a synaptic activity alteration occurs and consequently a neurodegeneration, as observed, for example, in Alzheimer's disease [96].

Gene GABRB3 encodes a protein named *Gamma-aminobutyric* acid receptor subunit beta-3 that is a ligand-gated ion channel and it is a receptor for gamma-aminobutyric acid (the major inhibitory neurotransmitter in the brain. This protein is expressed in different parts of the brain such as, the cerebral cortex, cerebellum, hippocampus, piriform cortex and thalamus, at different levels. Mutations in this gene and the protein low expression could lead to several neurologic diseases for example autism, epilepsy and Angelman syndrome [23, 90].

MECP2 gene encodes an essential epigenetic regulator protein in the brain development. This protein binds to methylated DNA, an epigenetic mark and can recruit transcriptional repressors to stop the gene expression or activate the expression of other genes [37].

SCN1A gene codes a sodium channel protein, neuronal voltage-gated sodium channel α-subunit type I, that regulates brain activity. Mutations on this gene could lead to several neurodevelopment diseases from several incident ranges, for example, simple febrile seizures to a severe epilepsies or locomotion's problems. Voltage-gated sodium channels are responsible for controlling electrical excitability and changes in this voltage can increase the sodium permeability and consequently does not allow the resting level of neuron [32].

CDKL5 is a kinase encoded by CDKL5 gene that adds phosphate groups to the DNA strand, to regulate the gene expression. This protein is mainly active in the brain and studies report that it is involved in the formation, grow and movement of the neurons but it also plays an important role in synapse (chemical signal connections between neurons) [44].

All of these coding genes seem to have important roles on the brain structure and brain activity, being also reported on the literature for some neurodegenerative diseases.

4.2.2.4. Cardiovascular diseases

Cardiovascular diseases cover vasculature, myocardium and the electrical circuit of heart [51]. We include these diseases in this work in the same category as anaemias. Heart failure is a progressive heath pathology which is a consequence of some injuries such as diabetes, cardiomyopathy, congenital disorders, hypertension, among others [34].

As shown in table 5, for cardiovascular diseases, were associated ACE, IL6, TNF, PPARG, ADIPOQ, SCN5A gene as the most important.

ACE gene encodes angiotensin-converting enzyme that cleaves proteins. This enzyme belongs to the renin-angiotensin system which regulates the blood pressure and the salt balance on body fluids, controlling the constriction of blood vessels and producing aldosterone hormone (regulates the absorption of salt and water in the kidneys), respectively [107].

A high expression of IL-6 (Interleukin-6) is associated with ventricular hypertrophy severity, i.e., myocardial injuries [50]. Studies have been reported that there is a relation between levels of TNF and IL-6 in the severity of heart failure. TNF and Interleukins are the most common molecules that act in inflammatory response mediating cardiac remodelling. The cardiac muscle maintains its tissue homeostasis resorting to innate and adaptative inflammatory responses [34]. This affirmation supports the fact of having two of the most relevant genes of autoimmune diseases in cardiac diseases.

Proliferator-activated receptor gamma (PPARG) is a gene encoding a nuclear hormone receptor that has an important role controlling the metabolism of lipids and glucose. This gene/protein is highly associated to Type 2 diabetes mellitus but also obesity and cancer, and depending on the gene mutation it could lead to the protection or risk of getting the disease. Genes of PPARG's family, PPARs genes, play important roles in several metabolic processes which makes them interesting as targets for metabolic syndrome treatments. However, when using agonist drugs to PPARs genes in

diabetes some side effects have been reported, so it is necessary to better understand the genomic pathogenesis of this disease [82].

Adiponectin gene (ADIPOQ) is reported as having a connection with some cardiac diseases once it influences the level and activity of adiponectin. Depending on the mutation, it could lead to risk of *Type 2 Diabetes Miellitus*, *Obesity* and *Coronary Artery Disease*. Adiponectin protein is produced mostly in adipose tissue but also in muscles and in the brain [63]. It regulates glucose levels and the breakdown of fat acids, however, the mechanism for these actions still needing more investigation [52].

SNC5 (sodium voltage-gated channel alpha subunit 5) gene is part of a gene family whose coded proteins by them, make part of sodium channels. These proteins are abundant on the heart and channels controlling the ions passing into cells. Actually, the main role of this channels is maintaining a normal heartbreak rhythm [16, 83].

All reported genes are highly associated to cardiovascular diseases. An interesting point is gene ADIPOQ which is also produced on brain and its mechanisms of action need to be more investigated, once could be a gene to focus on future works of drug repositioning.

4.2.2.5. Pulmonary Diseases

According to National Institute of Health (NIH), one of the most important medical research centers of the world, pulmonary diseases are type of diseases that affect lungs and other parts of respiratory system including asthma, lung cancer, pulmonary fibrosis, etc. According to table 5, the most common genes on complete genes are ADRB2, TGFB1, NOS3 genes and in short lists are IL5, IL4 and IL13 genes.

ADRB2 gene (Beta-2 Adrenergic Receptor) encodes a protein member of the G protein-couple receptor (GPCR) family and is highly associated to *Asthma* susceptivity. This protein is primary expressed on bronchial smooth muscle cells and is one of the responsible for the expansion of the small airways. Agonist of this proteins are used in asthma disease in order to do a bronchodilator therapy. Mutations on this gene are related to asthma severity, airways hypersensitive [58].

TGFB1 is a gene that encodes the protein TGF-β1 (transforming growth factor-β1) that belongs to a transforming growth factor beta superfamily of cytokines and has functions as controlling cell growth, differentiation, proliferation and induces the cell apoptosis. This gene is associated to chronic obstructive pulmonary disease that is characterized by airflow obstruction [21] but also it was reported as adolescent idiopathic scoliosis in Russian population and Chinese population [102]. Ultimately, according to a

study in 2012 by Barsova R.M. *et. al.,* [11] this gene is also related to Myocardial Infarction Susceptibility in Russian population. Besides of pulmonary diseases [21], TGFB1 gene could be associated with other diseases types such as cardiopathy [11] and bones diseases [102] which makes it a good candidate for therapies and drug repositioning.

NOS3 (Nitric oxid synthase 3) that encodes a protein nitric oxide (NO), appears to be related to pulmonary diseases, at least is present in 5/6 of our pulmonary diseases data. This protein mediates Vascular endothelial growth factor (VEGF) that mediates the angiogenesis (new blood vessels formation). NOS3 gene is highly associated to cardiac diseases but also to Alzheimer's disease. Using this consensus method this gene which could make it a new target for pulmonary diseases.

Interleukins are immunoregulatory cytokines that promote proliferation, differentiation of immune cells [84]. In our data, our main diseases are asthma and pulmonary hypertension that are pulmonary inflammatory diseases [46] so having interleukins as main genes on top 20 instead of others is totally possible.

4.2.3. Diseases group comparisons

The diseases comparison was made between diseases of the same type in pairs. We observed how many different pairs of diseases from same type share genes at least 1 gene on top 20 prioritized genes list and the mean of shared genes by all pairs that share more than 1 (this information is present in table 6).

Table 6. Tables that show the number of combinations of diseases that share genes and mean of shared genes by all diseases' combinations that share more than 1 per each disease type.

DISEASE TYPE	TOTAL DISEASES	NO. DISEASES PAIRS WHICH SHARE GENES	MEAN NO. SHARED GENES (>1)
CARDIO VASCULAR	40	133	4
PULMONARY	12	7	3
AUTOIMMUNE	19	139	3
NEUROLOGIC	38	119	2
CANCER	117	6069	2

For cardiac diseases, 133 combinations of diseases were obtained, sharing at least one gene. For pulmonary, autoimmune, neurologic and cancer were obtained 7,

139, 119 and 6069 combinations, respectively. The number of diseases for each type is not constant, therefore influencing the number of combinations. For better understanding, in cardiovascular diseases, 17% of all possible combinations of diseases pairs share at least 1 gene. The same happens for 11% in pulmonary diseases, 81% in autoimmune diseases, 17% in neurologic diseases and, finally, 89% in cancer.

Most disease combinations either share zero genes or one gene. This reveals that even between the same type, there are diseases with low similarity, the reason why table 6 shows the mean of shared genes by all disease combinations sharing more than one gene. The higher number of shared genes belongs to the cardiovascular diseases, with a mean of 4.

The most similar cardiac diseases are Beta Thalassemia Major and Beta Thalassemia sharing 16 genes, followed by Cardiac Arrest with Cardiac Arrythmia sharing 13 genes. Pseudohypoparathyroidism Type 1A with Albright Hereditary Osteodystrophy and Cardiac Arrythmia with Atrial Fibrillation share 11 genes and Parokysmal Supraventricular Tachycardia with Cardiac Arrest and Beta-Thalassemia with Sickle Cell Anemia are sharing 10 genes.

Among the most similar neurologic diseases are Epilepsy with Generalize Tonic-Clonic Seizures, Developmental Epileptic Encephalopathy and Dravet Syndrome, sharing at least 10 genes on the top 20 of prioritized genes.

In autoimmune diseases, the most similar are Psoriasis and Psoriatic Arthritis sharing 10 genes, followed by Lupus Erythematous with Rheumatoid Arthritis and with Secondary Progressive Multiple Sclerosis sharing 6 genes. Finally Psoriatic Arthritis with Chron Colitis share 5 genes.

The various Asthma types such as, atopic asthma, asthma and allergic asthma share 12 genes between them on pulmonary diseases group. No other pulmonary diseases are as similar as them.

Each type of diseases was combined with cancers data, but no similarity was observed. The maximum genes number shared were 2 and all of them were different depending on the disease.

In table 7 are reported all the genes shared by the combines diseases spoken above.

Table 7. Most similarity diseases by each type with respective shared genes.

TYPE	DISEASES COMBINATION			SHARED GENES
CARDIOVASCULAR	*Beta Thalassemia Major*	***VS***	*Beta Thalassemia*	*HFE, HBE1, TFRC, HBG1, HBA1, HBD, HBA2, GATA1, HBB, UGT1A1, VDR, HBG2, HAMP, COL1A1, BCL11A, KLF1*
	Cardiac Arrest	***VS***	*Cardiac Arrythmia*	*KCNJ2, KCND3, KCNH2, RYR2, SCN5A, HCN4, KCNE1, CACNA1C, DSP, KCNQ1*
	Pseudohypoparathyroidism Type 1A	***VS***	*Albright Hereditary Osteodystrophy*	*STX16, BDMR, GNG3, PTHR1, GNAS-AS1, PRKAR1A, GNAS, GNG4, PDE4D, PTH, GNG11*
	Cardiac Arrythmia	***VS***	*Atrial Fibrillation*	*KCNJ2, KCNE2, KCND3, KCNH2, RYR2, SCN5A, HCN4, NKX2-5, KCNE1, KCNE4, KCNQ1*
	Parokysmal Supraventricular Tachycardia	***VS***	*Cardiac Arrest*	*KCNJ2, PKP2, KCNH2, RYR2, SCN5A, CASQ2, HCN4, KCNE1, DSP, KCNQ1*
	Beta-Thalassemia	***VS***	*Sickle Cell Anemia*	*HBE1, G6PD, MYB, HBA1, HBA2, HBB, UGT1A1, HBG2, BCL11A, HBS1L*
NEUROLOGIC	*Epilepsy with Generalize Tonic-Clonic Seizures*	***VS***	*Developmental and Epileptic encephalopathy*	*PCDH19, SCN2A, KCNQ2, TBC1D24, CDKL5, SLC2A1, ARX, STXBP1, SCN8A, SCN1A*
	Epilepsy with Generalize Tonic-Clonic Seizures	***VS***	*Dravet Syndrome*	*EFHC1, PCDH19, SCN2A, SCN1B, KCNQ2, EPM2A, CSTB, GABRA1, SLC2A1, ARX, GABRG2, STXBP1, SCN8A, SCN1A*
	Developmental and Epileptic encephalopathy	***VS***	*Dravet Syndrome*	*PCDH19, SCN2A, KCNQ2, SLC2A1, ARX, STXBP1, SCN8A, SCN1A*

AUTOIMMUNE	*Psoriasis*	***VS***	*Psoriatic Arthritis*	*MICA, IL17A, HLA-B, IL12B, CCHCR1, IL23A, TRAF3IP2, KIR2DS1, TNF, IL23R*
	Rheumatoid Arthritis	***VS***	*Lupus Erythematous*	*IL18, PTPN22, CTLA4, HLA-DRB1, TNF, IL10*
	Lupus Erythematous	***VS***	*Secondary Progressive Multiple Sclerosis*	*CD40LG, IL18, FAS, CTLA4, HLA-DRB1, TNF*
	Psoriatic Arthritis	***VS***	*Chron Colitis*	*NOD2, IL12B, HLA-DRB1, TNF, IL23R*
PULMONARY	Allergic Asthma	***VS***	Atopic Asthma	*SPINK5, IL4R, CCL11, IL13, TSLP, TLR4, MS4A2, CCL5, HAVCR1, IL5, IL4, IL9*
	Atopic Asthma	***VS***	Asthma	*SPINK5, IL4R, CCL11, IL13, MS4A2, SCGB3A2, CCL5, CD14, NPSR1, ADAM33, IL5, IL4*
	Asthma	***VS***	Atopic Asthma	*SPINK5, IL4R, CCL11, IL13, MS4A2, SCGB3A2, CCL5, CD14, NPSR1, ADAM33, IL5, IL4*

4.3. Association Rules - Machine Learning Application

As explained in section 3.3, we made nine tests ranging the parameters (table 2) of *Apriori* Algorithm in three versions of the dataset (V1, V2, V3).

For each data version, V1, V2, V3, were obtained 44, 27 and 36 different rules, respectively. Rules presented in the next three tables (8, 9 and 10) were predicted, at least, in five tests (50%) in data V1, V2, V3 respectively.

Table 8. Rules predicted by at least 5 tests (50%) in data V1.

RULES	TIMES PREDICTED
Pembrolizumab → Bevacizumab	7
Bevacizumab → Paclitaxel	7
Pembrolizumab → Paclitaxel	7
Pembrolizumab → Ipilimumab	7
Bevacizumab → Temsirolimus	7
Bevacizumab → Gemcitabine	7

Pembrolizumab → Docetaxel	7
Bevacizumab → Ipilimumab	7
Paclitaxel → Temsirolimus	5
Sorafenib →Paclitaxel	5
Bevacizumab → Docetaxel	5
Cetuximab → Paclitaxel	5

Table 9. Rules predicted by at least 5 tests (50%) in data V2.

RULES	TIMES PREDICTED
Nivolumab → Ipilimumab	7
Pembrolizumab →Paclitaxel	7
Bevacizumab →Paclitaxel	7
Bevacizumab→ Pembrolizumab	7
Paclitaxel → Nivolumab	7
Pembrolizumab → Atezolizumab	7
Pembrolizumab → Nivolumab	7
Bevacizumab → Nivolumab	7

Table 10. Rules predicted by at least 5 tests (50%) in data V3.

RULES	TIMES PREDICTED
Pembrolizumab → Bevacizumab	7
Pembrolizumab →Paclitaxel	7
Bevacizumab →Paclitaxel	7
Pembrolizumab → Docetaxel	7
Nivolumab → Ipilimumab	7
Temsirolimus → Bevacizumab	7
Bevacizumab → Nivolumab	7
Pembrolizumab → Nivolumab	7
Pembrolizumab → Atezolizumab	7
Bevacizumab → Pembrolizumab	7
Paclitaxel → Atezolizumab	7
Paclitaxel → Nivolumab	7

By grouping all obtained rules in three datasets, 63 different rules were found, where 15 rules were predicted in all of them, 14 in two of them and 34 only in one dataset. However, the specific rules are the ones predicted in all versions, and table 11 shows those 15 rules.

Table 11.The fifteen specific rules predicted applying the Apriori Algorithm in all versions of data (V1, V2, V3).

SPECIFIC RULES
Gemcitabine → Bevacizumab
Gemcitabine → Paclitaxel

Gemcitabine → Docetaxel
Metformin → Bevacizumab
Pembrolizumab → Ipilimumab
Docetaxel → Paclitaxel
Metformin → Paclitaxel
Pembrolizumab → Paclitaxel
Bevacizumab → Paclitaxel
Bevacizumab → Ipilimumab
Cetuximab → Paclitaxel
Pembrolizumab → Docetaxel
Bevacizumab → Docetaxel
Pembrolizumab → Bevacizumab
Bevacizumab→Pegfilgrastim

Each one of these rules say that a drug could be applied for same disease as the previous. For example, Bevacizumab could be applied for the same disease or similar diseases as Gemcitabine, and so one.

In these rules are present nine different drugs that will be presented by a small description on the next point.

4.3.1. Drugs Description

Bevacizumab is a humanized monoclonal anti-vascular endothelial growth factor antibody approved by FDA for therapies. It binds and inhibits the vascular endothelial growth factor A (VEGFA) which has an important role in angiogenesis (physiologic process of creating new vessels from the pre-existing ones). Inhibiting angiogenesis is one of the best strategies to control the cancer metastasis, thus Bevacizumab is mainly suggested to treatment of several cancers (e.g., breast cancer, colorectal cancer, small cell lung cancer)[39].

Pembrolizumab and Ipilimumab are monoclonal antibodies also approved by the US FDA. These drugs have as targets the Programmed Cell Death Protein 1 (PDCD1) and the Cytotoxic T-lymphocyte antigen-4 (CTLA4) respectively, which are immune checkpoints. PDCD1 controls the immune responses preventing the overstimulation and avoiding the response to the self-antigens [71]. CTLA4 has a positive regulation of the regulatory T-Cells (which downregulate the immune response) thus blocking this protein increases the immune responses [61]. Ipilimumab was the first immune cell-targeting antibodies approved by FDA and is the standard treatment for cancers showing good long term results and protection against cancer recidivism [61, 79].

Paclitaxel and Docetaxel are small molecules approved by FDA which hyper-stabilize the microtubules in vivo, as opposed to other drugs that makes the depolymerization. Their function makes of mitotic cycle to stop and induces the death of tumor cells. This drug is indicated for several cancers therapy such as breast, ovarian and lung cancers but is also used prostate, endometrial and leukaemia [43, 98]. Paclitaxel and Docetaxel are among the best drugs to use in therapies of breast cancers early stages [41].

Metformin is a small molecule and is considered one of the most antidiabetic agents. This drug has been indicated for type 1 and 2 diabetes, paradiabetes, and gestational diabetes mellitus [31]. Reducing hyperglycaemia and the alleviation of its clinical symptoms are the most classical effects of metformin. This drug is a negative modulator that inhibits the mitochondrial complex I (NADH Dehydrogenase), preventing the production of mitochondrial ATP which leads to an AMP-activated protein kinase (AMPK) increasing. AMPK regulates the glucose metabolism, phosphorylating two isoforms of acetyl-CoA carboxylase enzyme. In the final of process this drug increases liver sensitivity to insulin and inhibits fat synthesis. Some recent studies reported Metformin has cadio and nephroprotective effects as well as reducing the neurodegenerative diseases incidence and the cancer progression [31].

Gemcitabine is a small molecule. It's a repositioning drug, initially tested as antiviral agent but currently it's mostly used in several cancer therapies such as ovarian, breast, lung and pancreatic, etc. This molecule has a good toxicity profile which makes it's useful from the young to the elderly people and an interesting molecule to be tested in combination with other cytotoxic agents. It mediates its antitumour effects by blocking the cell cycle by inhibiting DNA chain elongation and consequently promoting the apoptosis of malignant cells [42].

Cetuximab is a recombinant chimeric antibody that binds to Epidermal Growth Factor Receptor (EGFR), blocks phosphorylation and activation of receptor-associated kinases and avoid the cell metastasis, invasion proliferation and neovascularization. EGFR is a protein that regulates the development of epithelial tissue and homeostasis and it's overexpressed in tumoral cells. This drug has been investigated for colorectal cancer [104], non-small cell lung cancer [77] and unresectable squamous cell skin cancer [65].

Pegfilfrastim is a recombinant methionyl form of human granulocyte colony-stimulation factor (G-CSF) that increases the differentiation, proliferation and maturation of neutrophils, this drum mimics G-CSF binding to the same receptor (G-CSF receptor, protein coded by CSF3R gene) and has the same action. This drug is most indicated as an anticancer drug.

4.4. Drug Repositioning – New associations between Drugs-Diseases

According to the description of these rules, it is possible to conclude that almost of drugs presented on them are indicated for cancers. This fact could have two explanations: one is that many of created the drugs are focused in cancer once it is the most common cause of mortality and morbidity in the world [64]; the other explanation, is the unfeasible data because there are much more cancer related diseases than other types. The National Comprehensive Cancer Network (NCCN) reported that 50-70% drug repositioning discoveries or biologic therapies in the USA are approved for cancers [45]. This means that our data could be a good premise for future works.

Some of the obtained rules are already being tested, however not only with the repositioning but also with combination of drugs. For example, the combination of Gemcitabine with Paclitaxel has already been tested with success in metastic breast cancer diseases [29] as same as Pembrolizumab with Docetaxel on non-small cell lung cancer [9]. Others have been studied such as: Gemcitabine with Bevacizumab in matastic ovarian [67, 89] and pancreatic cancer treatment [53]; Gemcitabine combination with Docetaxel for treatment of sarcomas [40]; Cetuximab combination with Paclitaxel for neck squamous cell carcinoma treatment [26]; and finally, Bevacizumab with Docetaxel for several stages of ovarian cancer treatment [99].

Metformin is a drug mostly used for diabetes therapies, but this drug was also indicated for cancer and autoimmune diseases (neurodegenerative). This drug is associated with Bevacizumab which is also an interesting rule that besides cancer is also tested for respiratory diseases [73]. By interpreting the rule, we could say that Metformin can act in respiratory mechanisms and Bevacizumab in cardiovascular diseases mechanisms, and neurodegenerative or autoimmune diseases. In 2020, Yen *et al.* investigated the respiratory consequences of using Metformin in patients with coexistent Type 2 Diabetes Mellitus and Chronic Obstructive Pulmonary Disease and observed that it could lead to bacteria pneumonia or/and hyperglycaemia [105]. Furthermore, Bevacizumab had already been reported as having influence in Dementia and Parkinson diseases [86]. All these results are interesting to support our conclusions.

Paclitaxel is a drug that hyper-stabilizes the microtubules during the mitotic cycle which interrupts the cycle and induces the cell to apoptosis [98]. This drug as we can observe before, has been studied in combination with other drugs. There are two drugs that it is combined with Metformin and Bevacizumab. For the reasons presented before, we conclude that this drug could also have influence in cardiovascular or respiratory diseases. In 2017, Osman made a study which concluded that Paclitaxel has cardiotoxic

effects in patients with heart problems [72]. Also, Paclitaxel was studied as leading pneumonitis in patients with breast cancer [88]. Once again, we can say that this rule was well predicted by the model because it represents a mechanism that drugs share.

Pegfilfrastim is indicated as anticancer drug which target is CSF3R (Granulocyte colony-stimulating factor receptor). The protein is implied in neutrophiles proliferation and differentiation, which lead to an increasing of the immune response [95]. Autoimmune diseases appears when there is a deregulation in immune system that attack cells of the own organism. Increasing the immune response should not be the best method to fight the problem, so we conclude that this drug could have an important role in immune response, however, it should be applied in diseases that inhibit the immune activity. Pegfilfrastim and its biosimilars have been taking a high relevance in the market and their production and application (oncology, HIV, etc.) are expectable to increase during the next decade [2].

Chapter 5

Conclusions

This work approaches the first stages for drug repositioning application and also increases the libraries information to help future works. There are several studies that need information to compare, support and improve results, or use more complex machine learning models.

All objective delimited were concluded with success. The first objective, which comprises selection, collection and data treatment took a lot of time due to the data heterogeneity but also because the information is very spread out. In fact, the data diversity and complexity is one of the biggest problems for the models performance in DR [55].

When exploring the data, we observed that it corresponds to the actual market situation where small molecules are the predominant drugs used and developed [68]. Also, the literature supported the results of gene prioritization strategy for each disease type. All the genes presented on top 3 are highly associated with related diseases. There is no significant difference between top genes on complete prioritized list or in top 3, except on pulmonary diseases. This fact could be explained as our main pulmonary diseases are asthma and pulmonary hypertension related diseases which are pulmonary inflammatory diseases [46]. Therefore, having interleukins as main genes on top 20 instead of others is totally possible.

There are not a lot of similarity between diseases, either between each type of disease or with its related cancers. This was a premise that was supposed at the

beginning when 21 diseases were chosen. Due to this fact it is possible to conclude that focus only on gene prioritization methodology (Signature-based method) was not enough to describe diseases in order to compare them.

With the implementation of Association Rules, we obtained a set of fifteen rules. Some of them have been already described by literature and tested *in vivo*, mainly tested in combination, such as Gemcitabine with Paclitaxel in metastic breast cancer diseases [29] and Pembrolizumab with Docetaxel in non-small cell lung cancer [9], etc. This fact demonstrates the good quality of the data used and the model good performance.

Almost all the rules contained drugs indicated to cancer, with different targets and mechanisms of action.

We can highlight four drugs, which interactions could be interesting for future works: Metformin, Paclitaxel, Bevacizumab and Pegfilfrastim.

Metformin is mostly used for diabetes therapies (i.e., cardiovascular diseases) and both Paclitaxel and Bevacizumab are drugs indicated to cancer. However, according to the literature, we observed that Metformin has been studied for cancer and neurodegenerative diseases [31] and, also, Bevacizumab in respiratory diseases [73]. These drugs are present in some rules from which we conclude that: Metformin could have an impact in cancer and respiratory diseases; Paclitaxel could influence diseases as diabetes (or cardiovascular diseases) and respiratory diseases; and Bevacizumab could have influence in autoimmune and respiratory diseases.

Other important highlight was Pegfilfrastim that is indicated as anti-cancer drug. We observed that its target CSF3R (Granulocyte colony-stimulating factor receptor) is implied in neutrophiles proliferation and differentiation so we can also propose this drug for studies of deficit immune diseases.

We observed each one of our fifteen rules and we support our results with literature. All rules either have drugs with the same goal so they are being studied to be combined in order to get stronger and efficient treatments specially in cancer; or drugs has influence in same diseases mechanisms.

To conclude, as we can support the results with the literature, we can argue that our model is trustworthy for future works.

As said before this work should have more objectives but due the pandemic, time and the computational power, the goals changed significantly. The results were good but should be improved with more data, mainly describing diseases.

5.1. Future Work

Starting from the idea of diseases and drugs characterization and after the results, new approaches could be tested. There are several studies that use more complex machine learning models, such as neural networks, which need information to compare the results or even improve models.

Primarily, more information should be added to the data to improve the results, for example, add drug's structure and phenotypes of diseases.

Secondly, the application of different methodologies of similarity or distance, for example, Jaccard similarity, Tanimoto similarity, Euclidean distance or Hamming distance, should be done to calculate the similarity between drugs and diseases.

Finally, it would be also interesting the application of other machine learning models, for example, Inductive Logic Programming, a machine learning model that is capable of learning from a small set of data. Similarly, as AR, this model produces rules that are much easier to understand for Machine Learning layman.

Bibliography

1. *Apriori Algorithm in Python (Recommendation Engine)*. 03-03-2021]; Available from: https://deepak6446.medium.com/apriori-algorithm-in-python-recommendation-engine-5ba89bd1a6da.
2. *Autoimmune Disease Treatment Opportunities Provide Impetus to Pegfilgrastim Biosimilar Market*. 08-11-2021]; Available from: https://www.biosimilardevelopment.com/doc/autoimmune-disease-treatment-opportunities-pegfilgrastim-biosimilar-market-fact-mr-study-0001.
3. *Complete Guide to Association Rules*. 03-03-2021]; Available from: https://towardsdatascience.com/association-rules-2-aa9a77241654.
4. *Fda (Us Food and Drug Administration)*. 08-06-2021]; Available from: https://www.fda.gov/about-fda/cvm-offices/international-counterparts.
5. *Global Cancer Observatory: Cancer Today. Lyon, France: International Agency for Research on Cancer.* 10-06-2021]; Available from: https://gco.iarc.fr/today,.
6. Adasme, M.F., Parisi, D.*, et al. Structure-Based Drug Repositioning: Potential and Limits*. in *Seminars in cancer biology*. 2020. Elsevier.
7. Adasme, M.F., Parisi, D.*, et al.* **2020**. *Structure-Based Drug Repositioning Explains Ibrutinib as Vegfr2 Inhibitor. PloS one*. 15(5): p. e0233089.
8. Ananiadou, S. and Mcnaught, J. **2006**. *Text Mining for Biology and Biomedicine*. *Citeseer*.
9. Arrieta, O., Barrón, F.*, et al.* **2020**. *Efficacy and Safety of Pembrolizumab Plus Docetaxel Vs Docetaxel Alone in Patients with Previously Treated Advanced Non–Small Cell Lung Cancer: The Prolung Phase 2 Randomized Clinical Trial. JAMA oncology*. 6(6): p. 856.

10. Ashburn, T.T. and Thor, K.B. **2004**. *Drug Repositioning: Identifying and Developing New Uses for Existing Drugs. Nature reviews Drug discovery*. 3(8): p. 673.
11. Barsova, R., Titov, B.*, et al.* **2012**. *Contribution of the Tgfb1 Gene to Myocardial Infarction Susceptibility. Acta Naturae (англоязычная версия)*. 4(2 (13)).
12. Bisgin, H., Liu, Z.*, et al. Investigating Drug Repositioning Opportunities in Fda Drug Labels through Topic Modeling*. in *BMC bioinformatics*. 2012. Springer.
13. Bocci, V. **1991**. *Interleukins. Clinical pharmacokinetics*. 21(4): p. 274.
14. Boutorh, A., Pratanwanich, N.*, et al. Drug Repurposing by Optimizing Mining of Genes Target Association*. in *International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics*. 2014. Springer.
15. Bresso, E., Grisoni, R.*, et al.* **2013**. *Integrative Relational Machine-Learning for Understanding Drug Side-Effect Profiles. BMC bioinformatics*. 14(1): p. 207.
16. Brugada, P. **2016**. *Brugada Syndrome: More Than 20 Years of Scientific Excitement. Journal of cardiology*. 67(3): p. 215.
17. Byers, L.A., Diao, L.*, et al.* **2013**. *An Epithelial–Mesenchymal Transition Gene Signature Predicts Resistance to Egfr and Pi3k Inhibitors and Identifies Axl as a Therapeutic Target for Overcoming Egfr Inhibitor Resistance. Clinical cancer research*. 19(1): p. 279.
18. Cabrera-Andrade, A., López-Cortés, A.*, et al.* **2020**. *Gene Prioritization through Consensus Strategy, Enrichment Methodologies Analysis, and Networking for Osteosarcoma Pathogenesis. International journal of molecular sciences*. 21(3): p. 1053.
19. Carabotti, M., Scirocco, A.*, et al.* **2015**. *The Gut-Brain Axis: Interactions between Enteric Microbiota, Central and Enteric Nervous Systems. Annals of gastroenterology: quarterly publication of the Hellenic Society of Gastroenterology*. 28(2): p. 203.
20. Ceccarelli, F., Agmon-Levin, N.*, et al.* **2017**. *Genetic Factors of Autoimmune Diseases 2017*, *Hindawi*.
21. Celedon, J.C., Lange, C.*, et al.* **2004**. *The Transforming Growth Factor-B1 (Tgfb1) Gene Is Associated with Chronic Obstructive Pulmonary Disease (Copd). Human molecular genetics*. 13(15): p. 1649.
22. Chaudhari, R., Tan, Z.*, et al.* **2017**. *Computational Polypharmacology: A New Paradigm for Drug Discovery. Expert opinion on drug discovery*. 12(3): p. 279.
23. Chen, C.-H., Huang, C.-C.*, et al.* **2014**. *Genetic Analysis of Gabrb3 as a Candidate Gene of Autism Spectrum Disorders. Molecular autism*. 5(1): p. 1.

24. Chen, Y., Wang, W.*, et al.* **2011**. *In Silico Gene Prioritization by Integrating Multiple Data Sources. PloS one*. 6(6): p. e21137.
25. Cheng, D., Knox, C.*, et al.* **2008**. *Polysearch: A Web-Based Text Mining System for Extracting Relationships between Human Diseases, Genes, Mutations, Drugs and Metabolites. Nucleic acids research*. 36(suppl_2): p. W399.
26. Chevalier, T., Daste, A.*, et al.* **2021**. *Cetuximab Combined with Paclitaxel or Paclitaxel Alone for Patients with Recurrent or Metastatic Head and Neck Squamous Cell Carcinoma Progressing after Extreme. Cancer Medicine*.
27. Cho, J.H. and Feldman, M. **2015**. *Heterogeneity of Autoimmune Diseases: Pathophysiologic Insights from Genetics and Implications for New Therapies. Nature medicine*. 21(7): p. 730.
28. Cicenas, J., Zalyte, E.*, et al.* **2018**. *Kinases and Cancer. Cancers (Basel)*. 10(3).
29. Colomer, R. **2005**. *Gemcitabine in Combination with Paclitaxel for the Treatment of Metastatic Breast Cancer. Women's Health*. 1(3): p. 323.
30. Cruz-Monteagudo, M., Borges, F.*, et al.* **2016**. *Efficient and Biologically Relevant Consensus Strategy for Parkinson's Disease Gene Prioritization. BMC medical genomics*. 9(1): p. 1.
31. Drzewoski, J. and Hanefeld, M. **2021**. *The Current and Potential Therapeutic Use of Metformin—the Good Old Drug. Pharmaceuticals*. 14(2): p. 122.
32. Escayg, A. and Goldin, A.L. **2010**. *Sodium Channel Scn1a and Epilepsy: Mutations and Mechanisms. Epilepsia*. 51(9): p. 1650.
33. Fontaine, J.-F., Priller, F.*, et al.* **2011**. *Genie: Literature-Based Gene Prioritization at Multi Genomic Scale. Nucleic acids research*. 39(suppl_2): p. W455.
34. Fragoso, J.M. **2014**. *Tumor Necrosis Factor Alpha (Tnf-A) in Autoimmune Diseases (Aids): Molecular Biology and Genetics. Gaceta medica de Mexico*. 150(4): p. 334.
35. Gholami, A.M., Hahne, H.*, et al.* **2013**. *Global Proteome Analysis of the Nci-60 Cell Line Panel. Cell reports*. 4(3): p. 609.
36. Gill, N., Singh, S.*, et al.* **2014**. *Computational Disease Gene Prioritization: An Appraisal. Journal of Computational Biology*. 21(6): p. 456.
37. Gonzales, M.L. and Lasalle, J.M. **2010**. *The Role of Mecp2 in Brain Development and Neurodevelopmental Disorders. Current psychiatry reports*. 12(2): p. 127.
38. Guney, E., Garcia-Garcia, J.*, et al.* **2014**. *Guildify: A Web Server for Phenotypic Characterization of Genes through Biological Data Integration and Network-Based Prioritization Algorithms. Bioinformatics*. 30(12): p. 1789.

39. Han, K., Peyret, T.*, et al.* **2016**. *Population Pharmacokinetics of Bevacizumab in Cancer Patients with External Validation. Cancer chemotherapy and pharmacology*. 78(2): p. 341.
40. Hara, H., Kawamoto, T.*, et al.* **2019**. *Gemcitabine and Docetaxel Combination Chemotherapy for Advanced Bone and Soft Tissue Sarcomas: Protocol for an Open-Label, Non-Randomised, Phase 2 Study. BMC cancer*. 19(1): p. 1.
41. Ho, M.Y. and Mackey, J.R. **2014**. *Presentation and Management of Docetaxel-Related Adverse Effects in Patients with Breast Cancer. Cancer management and research*. 6: p. 253.
42. Hristovski, D., Peterlin, B.*, et al.* **2005**. *Using Literature-Based Discovery to Identify Disease Candidate Genes. International journal of medical informatics*. 74(2-4): p. 289.
43. Imran, M., Saleem, S.*, et al.* **2020**. *Docetaxel: An Update on Its Molecular Mechanisms, Therapeutic Trajectory and Nanotechnology in the Treatment of Breast, Lung and Prostate Cancer. Journal of Drug Delivery Science and Technology*. p. 101959.
44. Jakimiec, M., Paprocka, J.*, et al.* **2020**. *Cdkl5 Deficiency Disorder—a Complex Epileptic Encephalopathy. Brain sciences*. 10(2): p. 107.
45. Jin, G. and Wong, S.T. **2014**. *Toward Better Drug Repositioning: Prioritizing and Integrating Existing Methods into Efficient Pipelines. Drug discovery today*. 19(5): p. 637.
46. Joshi, B.H., Hogaboam, C.*, et al.* **2006**. *Role of Interleukin-13 in Cancer, Pulmonary Fibrosis, and Other Th2-Type Diseases. Vitamins & Hormones*. 74: p. 479.
47. Jourdan, J.P., Bureau, R.*, et al.* **2020**. *Drug Repositioning: A Brief Overview. Journal of Pharmacy and Pharmacology*. 72(9): p. 1145.
48. Jourquin, J., Duncan, D.*, et al.* **2012**. *Glad4u: Deriving and Prioritizing Gene Lists from Pubmed Literature. BMC genomics*. 13(8): p. 1.
49. Jung, J. and Lee, D. **2013**. *Inferring Disease Association Using Clinical Factors in a Combinatorial Manner and Their Use in Drug Repositioning. Bioinformatics*. 29(16): p. 2017.
50. Kanda, T. and Takahashi, T. **2004**. *Interleukin-6 and Cardiovascular Diseases. Japanese heart journal*. 45(2): p. 183.
51. Kathiresan, S. and Srivastava, D. **2012**. *Genetics of Human Cardiovascular Disease. Cell*. 148(6): p. 1242.

52. Khabour, O.F., Alomari, M.A.*, et al.* **2018**. *The Relationship of Adiponectin Level and Adipoq Gene Variants with Bmi among Young Adult Women. Dermato-endocrinology*. 10(1): p. e1477902.
53. Kindler, H.L., Friberg, G.*, et al.* **2005**. *Phase Ii Trial of Bevacizumab Plus Gemcitabine in Patients with Advanced Pancreatic Cancer. Journal of Clinical Oncology*. 23(31): p. 8033.
54. Kola, I. and Landis, J. **2004**. *Can the Pharmaceutical Industry Reduce Attrition Rates? Nature reviews Drug discovery*. 3(8): p. 711.
55. Koromina, M., Pandi, M.-T.*, et al.* **2019**. *Rethinking Drug Repositioning and Development with Artificial Intelligence, Machine Learning, and Omics. Omics: a journal of integrative biology*. 23(11): p. 539.
56. Kuo, M., Kushniruk, A.*, et al.* **2009**. *Application of the Apriori Algorithm for Adverse Drug Reaction Detection*, in *Detection and Prevention of Adverse Drug EventsIOS Press*. p. 95-101.
57. Lee, E.Y. and Muller, W.J. **2010**. *Oncogenes and Tumor Suppressor Genes. Cold Spring Harbor perspectives in biology*. 2(10): p. a003236.
58. Liang, S.-Q., Chen, X.-L.*, et al.* **2014**. *Beta-2 Adrenergic Receptor (Adrb2) Gene Polymorphisms and the Risk of Asthma: A Meta-Analysis of Case-Control Studies. PloS one*. 9(8): p. e104488.
59. Liekens, A.M., De Knijf, J.*, et al.* **2011**. *Biograph: Unsupervised Biomedical Knowledge Discovery Via Automated Hypothesis Generation. Genome biology*. 12(6): p. 1.
60. Little, J., Barakat-Haddad, C.*, et al.* **2017**. *Genetic Variation Associated with the Occurrence and Progression of Neurological Disorders. Neurotoxicology*. 61: p. 243.
61. Liu, Y. and Zheng, P. **2018**. *How Does an Anti-Ctla-4 Antibody Promote Cancer Immunity? Trends in immunology*. 39(12): p. 953.
62. López-Cortés, A., Paz-Y-Miño, C.*, et al.* **2018**. *Gene Prioritization, Communality Analysis, Networking and Metabolic Integrated Pathway to Better Understand Breast Cancer Pathogenesis. Scientific reports*. 8(1): p. 1.
63. Martinez-Huenchullan, S.F., Tam, C.S.*, et al.* **2020**. *Skeletal Muscle Adiponectin Induction in Obesity and Exercise. Metabolism*. 102: p. 154008.
64. Mbele, M., Hull, R.*, et al.* **2017**. *African Medicinal Plants and Their Derivatives: Current Efforts Towards Potential Anti-Cancer Drugs. Experimental and molecular pathology*. 103(2): p. 121.

65. Montaudié, H., Viotti, J.*, et al.* **2020**. *Cetuximab Is Efficient and Safe in Patients with Advanced Cutaneous Squamous Cell Carcinoma: A Retrospective, Multicentre Study. Oncotarget*. 11(4): p. 378.
66. Morgan, S., Grootendorst, P.*, et al.* **2011**. *The Cost of Drug Development: A Systematic Review. Health policy*. 100(1): p. 4.
67. Nagao, S., Kogiku, A.*, et al.* **2020**. *A Phase Ii Study of the Combination Chemotherapy of Bevacizumab and Gemcitabine in Women with Platinum-Resistant Recurrent Epithelial Ovarian, Primary Peritoneal, or Fallopian Tube Cancer. Journal of ovarian research*. 13(1): p. 1.
68. Ngo, H.X. and Garneau-Tsodikova, S. **2018**. *What Are the Drugs of the Future? MedChemComm*. 9(5): p. 757.
69. Nowak-Sliwinska, P., Scapozza, L.*, et al.* **2019**. *Drug Repurposing in Oncology: Compounds, Pathways, Phenotypes and Computational Approaches for Colorectal Cancer. Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*. 1871(2): p. 434.
70. Oh, M., Ahn, J.*, et al.* **2014**. *A Network-Based Classification Model for Deriving Novel Drug-Disease Associations and Assessing Their Molecular Actions. PloS one*. 9(10): p. e111668.
71. Okazaki, T. and Honjo, T. **2007**. *Pd-1 and Pd-1 Ligands: From Discovery to Clinical Application. International immunology*. 19(7): p. 813.
72. Osman, M. and Elkady, M. **2017**. *A Prospective Study to Evaluate the Effect of Paclitaxel on Cardiac Ejection Fraction. Breast Care*. 12(4): p. 255.
73. Pang, J., Xu, F.*, et al.* **2021**. *Efficacy and Tolerability of Bevacizumab in Patients with Severe Covid-19. Nature communications*. 12(1): p. 1.
74. Paul, S.M., Mytelka, D.S.*, et al.* **2010**. *How to Improve R&D Productivity: The Pharmaceutical Industry's Grand Challenge. Nature reviews Drug discovery*. 9(3): p. 203.
75. Peraldo-Neia, C., Migliardi, G.*, et al.* **2011**. *Epidermal Growth Factor Receptor (Egfr) Mutation Analysis, Gene Expression Profiling and Egfr Protein Expression in Primary Prostate Cancer. BMC cancer*. 11(1): p. 1.
76. Piñero, J., Queralt-Rosinach, N.*, et al.* **2015**. *Disgenet: A Discovery Platform for the Dynamical Exploration of Human Diseases and Their Genes. Database*. 2015.
77. Pirker, R., Pereira, J.R.*, et al.* **2009**. *Cetuximab Plus Chemotherapy in Patients with Advanced Non-Small-Cell Lung Cancer (Flex): An Open-Label Randomised Phase Iii Trial. The Lancet*. 373(9674): p. 1525.

78. Rastegar-Mojarad, M., Elayavilli, R.K.*, et al.* **2015**. *A New Method for Prioritizing Drug Repositioning Candidates Extracted by Literature-Based Discovery*. in *2015 IEEE international conference on Bioinformatics and Biomedicine (BIBM)*. IEEE.
79. Robert, C., Schachter, J.*, et al.* **2015**. *Pembrolizumab Versus Ipilimumab in Advanced Melanoma. New England Journal of Medicine*. 372(26): p. 2521.
80. Ryan, K.M., Phillips, A.C.*, et al.* **2001**. *Regulation and Function of the P53 Tumor Suppressor Protein. Current opinion in cell biology*. 13(3): p. 332.
81. Sansom, D.M. and Walker, L.S. **2006**. *The Role of Cd28 and Cytotoxic T-Lymphocyte Antigen-4 (Ctla-4) in Regulatory T-Cell Biology. Immunological reviews*. 212(1): p. 131.
82. Sarhangi, N., Sharifi, F.*, et al.* **2020**. *Pparg (Pro12ala) Genetic Variant and Risk of T2dm: A Systematic Review and Meta-Analysis. Scientific reports*. 10(1): p. 1.
83. Schulze-Bahr, E., Eckardt, L.*, et al.* **2003**. *Sodium Channel Gene (Scn5a) Mutations in 44 Index Patients with Brugada Syndrome: Different Incidences in Familial and Sporadic Disease. Human mutation*. 21(6): p. 651.
84. She, Y.X., Yu, Q.Y.*, et al.* **2021**. *Role of Interleukins in the Pathogenesis of Pulmonary Fibrosis. Cell Death Discovery*. 7(1): p. 1.
85. Simmonds, M. and Gough, S. **2007**. *The Hla Region and Autoimmune Disease: Associations and Mechanisms of Action. Current genomics*. 8(7): p. 453.
86. Sultana, J., Scondotto, G.*, et al.* **2020**. *Intravitreal Anti-Vegf Drugs and Signals of Dementia and Parkinson-Like Events: Analysis of the Vigibase Database of Spontaneous Reports. Frontiers in pharmacology*. 11: p. 315.
87. Swamidass, S.J. **2011**. *Mining Small-Molecule Screens to Repurpose Drugs. Briefings in bioinformatics*. 12(4): p. 327.
88. Taghian, A.G., Assaad, S.I.*, et al.* **2001**. *Risk of Pneumonitis in Breast Cancer Patients Treated with Radiation Therapy and Combination Chemotherapy with Paclitaxel. Journal of the National Cancer Institute*. 93(23): p. 1806.
89. Takasaki, K., Miyamoto, M.*, et al.* **2018**. *Addition of Bevacizumab to Gemcitabine for Platinum-Resistant Recurrent Ovarian Cancer: A Retrospective Analysis. Cancer chemotherapy and pharmacology*. 81(5): p. 809.
90. Tanaka, M., Delorey, T.M.*, et al.* **2012**. *Gabrb3, Epilepsy, and Neurodevelopment.*
91. Tejera, E., Cruz-Monteagudo, M.*, et al.* **2017**. *Consensus Strategy in Genes Prioritization and Combined Bioinformatics Analysis for Preeclampsia Pathogenesis. BMC medical genomics*. 10(1): p. 50.
92. Tranchevent, L.-C., Capdevila, F.B.*, et al.* **2011**. *A Guide to Web Tools to Prioritize Candidate Genes. Briefings in bioinformatics*. 12(1): p. 22.

93. Vatansever, S., Schlessinger, A.*, et al.* **2021**. *Artificial Intelligence and Machine Learning-Aided Drug Discovery in Central Nervous System Diseases: State-of-the-Arts and Future Directions. Medicinal Research Reviews*. 41(3): p. 1427.
94. Vogt, P. **1993**. *Cancer Genes. Western journal of medicine*. 158(3): p. 273.
95. Wang, X., Qiu, L.*, et al.* **2018**. *Understanding the Multifaceted Role of Neutrophils in Cancer and Autoimmune Diseases. Frontiers in immunology*. 9: p. 2456.
96. Wang, Y., Yang, P.-l.*, et al.* **2008**. *Potassium Channels: Possible New Therapeutic Targets in Parkinson's Disease. Medical hypotheses*. 71(4): p. 546.
97. Watanabe, H., Nagata, E.*, et al.* **2000**. *Disruption of the Epilepsy Kcnq2 Gene Results in Neural Hyperexcitability. Journal of neurochemistry*. 75(1): p. 28.
98. Weaver, B.A. **2014**. *How Taxol/Paclitaxel Kills Cancer Cells. Molecular biology of the cell*. 25(18): p. 2677.
99. Wenham, R.M., Lapolla, J.*, et al.* **2013**. *A Phase Ii Trial of Docetaxel and Bevacizumab in Recurrent Ovarian Cancer within 12 Months of Prior Platinum-Based Chemotherapy. Gynecologic oncology*. 130(1): p. 19.
100. Wu, X., Jiang, R.*, et al.* **2008**. *Network-Based Global Inference of Human Disease Genes. Molecular systems biology*. 4(1): p. 189.
101. Xie, B., Agam, G.*, et al.* **2015**. *Disease Gene Prioritization Using Network and Feature. Journal of Computational Biology*. 22(4): p. 313.
102. Xu, L., Sun, W.*, et al.* **2016**. *The Tgfb1 Gene Is Associated with Curve Severity but Not with the Development of Adolescent Idiopathic Scoliosis: A Replication Study in the Chinese Population. BMC musculoskeletal disorders*. 17(1): p. 1.
103. Yang, H., Robinson, P.N.*, et al.* **2015**. *Phenolyzer: Phenotype-Based Prioritization of Candidate Genes for Human Diseases. Nature methods*. 12(9): p. 841.
104. Yazdi, M.H., Faramarzi, M.A.*, et al.* **2015**. *A Comprehensive Review of Clinical Trials on Egfr Inhibitors Such as Cetuximab and Panitumumab as Monotherapy and in Combination for Treatment of Metastatic Colorectal Cancer. Avicenna journal of medical biotechnology*. 7(4): p. 134.
105. Yen, F.-S., Wei, J.C.-C.*, et al.* **2020**. *Respiratory Outcomes of Metformin Use in Patients with Type 2 Diabetes and Chronic Obstructive Pulmonary Disease. Scientific reports*. 10(1): p. 1.
106. Yue, P., Melamud, E.*, et al.* **2006**. *Snps3d: Candidate Gene and Snp Selection for Association Studies. BMC bioinformatics*. 7(1): p. 1.
107. Zhang, Z., Xu, G.*, et al.* **2012**. *Angiotensin-Converting Enzyme Insertion/Deletion Polymorphism Contributes to Ischemic Stroke Risk: A Meta-Analysis of 50 Case-Control Studies.*

108. Zhao, B.-W., You, Z.-H.*, et al.* **2021**. *A Novel Method to Predict Drug-Target Interactions Based on Large-Scale Graph Representation Learning. Cancers*. 13(9): p. 2111.

Annexes

I- List of Diseases Names (226 diseases)

0 Alzheimer

1 Congenital Dyserythropoietic Anemia Type I

2 Hemolytic Anemia

3 Refractory Anemia with Ringed Sideroblasts

4 Hemolytic-Uremic Syndrome

5 Anemia

6 Pure Red-Cell Aplasia

7 Congenital Amegakaryocytic Thrombocytopenia

8 Beta-Thalassemia Major

9 Paroxysmal Nocturnal Hemoglobinuria

10 Anemia, Hemolytic, Autoimmune

11 Beta-Thalassemia

12 Primary Myelofibrosis

13 Sickle Cell Anemia

14 Fanconi Anemia

15 Status Asthmaticus

16 Allergic Asthma

17 Atopic Asthma

18 Asthma

19 Autism

20 Pituitary Cancer

21 Acth-Producing Pituitary Gland Carcinoma

22 Childhood Brain Stem Glioma

23 Brainstem Cancer

24 Pituitary Adenocarcinoma ()

25 Pineoblastoma

26 Brain Glioma

27 Olfactory Neuroblastoma

28 Brain Stem Glioma

29 Gliosarcoma

30 Brain Cancer

31 Diffuse Intrinsic Pontine Glioma

32 Salivary Duct Carcinoma

33 Invasive Lobular Carcinoma

34 Female Breast Carcinoma

35 Invasive Breast Carcinoma

36 Luminal B Breast Carcinoma

37 Atypical Lobular Breast Hyperplasia

38 Hereditary Breast and Ovarian Cancer Syndrome

39 Breast Adenocarcinoma

40 Inflammatory Breast Carcinoma

41 Breast Ductal Adenocarcinoma

42 Triple-Negative Breast Cancer

43 Invasive Breast Ductal Carcinoma

44 Male Breast Carcinoma

45 Breast Carcinoma In Situ

46 Breast Carcinoma

47 Breast Cancer

48 Breast Ductal Carcinoma in Situ

49 Paroxysmal Supraventricular Tachycardia

50 Supraventricular Tachycardia

51 Atrial Flutter

52 Ventricular Tachycardia

53 Sudden Cardiac Arrest

54 Ventricular Fibrillation

55 Atrial Fibrillation

56 Cardiac Arrhythmia

57 Cardiac Arrest

58 Crohn Colitis

59 Crohn

60 Gestational Diabetes

61 Latent Autoimmune Diabetes in Adults

62 Prediabetes Syndrome

63 Type I Diabetes Mellitus

64 Stiff-Person Syndrome

65 Diabetes Mellitus

66 Type II Diabetes Mellitus

67 Rolandic Epilepsy

68 Lennox-Gastaut Syndrome

69 Epilepsia Partialis Continua

70 Epilepsy with Generalized Tonic-Clonic Seizures

71 Dravet Syndrome

72 Complex Partial Epilepsy

73 Status Epilepticus

74 Zellweger Syndrome

75 West Syndrome

76 Angelman Syndrome

77 Klinefelter Syndrome

78 Juvenile Neuronal Ceroid Lipofuscinosis

79 Partial Epilepsy

80 Japanese Encephalitis

81 Landau-Kleffner Syndrome

82 Granulomatosis with Polyangiitis

83 Childhood Absence Epilepsy

84 Reflex Epilepsy

85 Developmental and Epileptic Encephalopathy

86 Churg-Strauss Syndrome

87 Adrenomyeloneuropathy

88 Fragile X Syndrome

89 Rett Syndrome

90 Wilson

91 Gaucher Type 2

92 X-Linked Adrenoleukodystrophy

93 Epilepsy

94 Gangliosidosis

95 Behcet Syndrome

96 Colon Mucinous Adenocarcinoma

97 Appendix Cancer

98 Small Intestinal Adenocarcinoma

99 Anal Squamous Cell Carcinoma

100 Ampulla of Vater Adenocarcinoma

101 Anal Carcinoma

102 Colon Carcinoma

103 Small Intestinal Neuroendocrine Tumor G1

104 Colon Adenocarcinoma

105 Rectal Carcinoma

106 Colorectal Cancer

107 Anus Cancer

108 Gastric Intestinal Type Adenocarcinoma

109 Small Intestine Lymphoma

110 Appendix Carcinoma

111 Colorectal Adenocarcinoma

112 Lynch Syndrome

113 Small Intestine Cancer

114 Rectum Cancer

115 Malignant Colon Neoplasm

116 Metastatic Colorectal Cancer

117 Small Intestine Carcinoma

118 Colorectal Carcinoma

119 Bronchoalveolar Adenocarcinoma

120 Lung Carcinoid Tumor

121 Large Cell Lung Carcinoma

122 Pulmonary Neuroendocrine Tumor

123 Small Cell Lung Carcinoma

124 Lung Carcinoma

125 Adenosquamous Lung Carcinoma

126 Lung Cancer

127 Squamous Cell Lung Carcinoma

128 Non-Small Cell Lung Carcinoma

129 Lung Adenocarcinoma

130 Lupus Erythematosus

131 Systemic Lupus Erythematosus

132 Cutaneous Lupus Erythematosus

133 Secondary Progressive Multiple Sclerosis

134 Primary Progressive Multiple Sclerosis

135 Relapsing-Remitting Multiple Sclerosis

136 Chronic Progressive Multiple Sclerosis

137 Multiple Sclerosis

138 Neuromyelitis Optica

139 Bardet-Biedl Syndrome

140 Pseudohypoparathyroidism Type 1a

141 Morbid Obesity

142 Albright Hereditary Osteodystrophy

143 Rubinstein-Taybi Syndrome

144 Prader-Willi Syndrome

145 Obesity

146 Postencephalitic Parkinson

147 Secondary Parkinson

148 Parkinson

149 Prostate Small Cell Carcinoma

150 Metastatic Prostate Cancer

151 Prostate Adenocarcinoma

152 Prostate Carcinoma

153 Prostate Cancer

154 Generalized Pustular Psoriasis

155 Parapsoriasis

156 Psoriasis

157 Psoriasis Vulgaris

158 Pulmonary Venoocclusive

159 Pulmonary Arterial Hypertension

160 Pulmonary Hypertension

161 Ankylosing Spondylitis

162 Rheumatoid Arthritis

163 Psoriatic Arthritis

164 Schizophrenia

165 Pancreatic Adenosquamous Carcinoma

166 Extramammary Paget Disease

167 Dermatofibrosarcoma Protuberans

168 Lentigo Maligna Melanoma

169 Desmoplastic Melanoma

170 Endometrial Adenosquamous Carcinoma

171 Human Papillomavirus-Related Squamous Cell Carcinoma

172 Acral Lentiginous Melanoma

173 Mucosal Melanoma

174 Esophageal Squamous Cell Carcinoma

175 Gorlin Syndrome

176 Vulvar Squamous Cell Carcinoma

177 Lip and Oral Cavity Squamous Cell Carcinoma

178 Squamous Cell Carcinoma Of Penis

179 Oropharynx Squamous Cell Carcinoma

180 Extranodal Nasal Nk/T Cell Lymphoma

181 Ocular Melanoma

182 Choroidal Melanoma

183 Oral Squamous Cell Carcinoma

184 Uveal Melanoma

185 Endometrial Squamous Cell Carcinoma

186 Skin Squamous Cell Carcinoma

187 Salivary Gland Squamous Cell Carcinoma

188 Verrucous Carcinoma

189 Bladder Squamous Cell Carcinoma

190 Cutaneous Melanoma

191 Adult T-Cell Leukemia/ Lymphoma

192 Metastatic Melanoma

193 Sezary

194 Mycosis Fungoides

195 Nasal Cavity and Paranasal Sinus Squamous Cell Carcinoma

196 Melanoma

197 Squamous Cell Carcinoma

198 Head and Neck Squamous Cell Carcinoma

199 Non-Melanoma Skin Carcinoma

200 Basal Cell Carcinoma

201 Laryngeal Squamous Cell Carcinoma

202 Skin Cancer

203 Cutaneous Squamous Cell Carcinoma

204 Cutaneous T-Cell Lymphoma

205 Cervical Squamous Cell Carcinoma

206 Endometrial Stromal Sarcoma

207 Endometrial Clear Cell Adenocarcinoma

208 Uterine Cervix Carcinoma In Situ

209 Uterine Leiomyosarcoma

210 Uterine Carcinosarcoma

211 Endometrial Mixed Adenocarcinoma

212 Uterine Sarcoma

213 Endometrial Undifferentiated Carcinoma

214 Cervical Adenosarcoma

215 Cervical Cancer

216 Endometrial Serous Adenocarcinoma

217 Endometrium Adenocarcinoma

218 Endometrial Mucinous Adenocarcinoma

219 Cervical Carcinosarcoma

220 Cervical Adenocarcinoma

221 Cervical Carcinoma

222 Uterine Cancer

223 Endometrial Carcinoma

224 Cervical Intraepithelial Neoplasia Grade 2/3

225 Endometrial Cancer

II- Drugs described by Type, Action Type and Activity

Table containing all drugs of our data characterized by Type, Action Type, Activity.

Drug	Type	Action type	Activity
Abatacept	Protein	Inhibitor	Negative modulator
Abiraterone	Small molecule	Antagonist	Negative modulator
Abiraterone	Small molecule	Inhibitor	Negative modulator
Acamprosate	Small molecule	Antagonist	Negative modulator
Acamprosate	Small molecule	Positive modulator	Positive modulator
Acetaminophen	Small molecule	Inhibitor	Negative modulator
Acetaminophen	Small molecule	Opener	Positive modulator
Acitretin	Small molecule	Agonist	Positive modulator
Afatinib	Small molecule	Inhibitor	Negative modulator
Aflibercept	Protein	Inhibitor	Negative modulator
Aldesleukin	Protein	Agonist	Positive modulator
Alitretinoin	Small molecule	Agonist	Positive modulator
Alprazolam	Small molecule	Positive allosteric modulator	Positive modulator
Alvocidib	Small molecule	Inhibitor	Negative modulator
Amantadine	Small molecule	Antagonist	Negative modulator
Amifampridine	Small molecule	Blocker	Negative modulator
Amiselimod	Small molecule	Modulator	Other
Amisulpride	Small molecule	Antagonist	Negative modulator
Amlodipine	Small molecule	Blocker	Negative modulator
Anlotinib	Small molecule	Inhibitor	Negative modulator
Apitolisib	Small molecule	Inhibitor	Negative modulator
Apremilast	Small molecule	Inhibitor	Negative modulator
Aripiprazole	Small molecule	Antagonist	Negative modulator
Aripiprazole	Small molecule	Partial agonist	Positive modulator
Asm-024	Small molecule	Antagonist	Negative modulator
Asm-024	Small molecule	Modulator	Other
Aspirin	Small molecule	Inhibitor	Negative modulator
Atorvastatin	Small molecule	Inhibitor	Negative modulator
Axitinib	Small molecule	Inhibitor	Negative modulator
Azacitidine	Small molecule	Inhibitor	Negative modulator
Azd-4547	Small molecule	Inhibitor	Negative modulator
Azd1305	Small molecule	Blocker	Negative modulator
Becatecarin	Small molecule	Inhibitor	Negative modulator
Belinostat	Small molecule	Inhibitor	Negative modulator
Bevacizumab	Antibody	Inhibitor	Negative modulator
Bexarotene	Small molecule	Agonist	Positive modulator
Binimetinib	Small molecule	Inhibitor	Negative modulator
Birabresib	Small molecule	Inhibitor	Negative modulator
Bms-690514	Small molecule	Inhibitor	Negative modulator
Bortezomib	Small molecule	Inhibitor	Negative modulator

Brentuximab vedotin	Antibody	Inhibitor	Negative modulator
Brexanolone	Small molecule	Positive allosteric modulator	Positive modulator
Brivanib	Small molecule	Inhibitor	Negative modulator
Buparlisib	Small molecule	Inhibitor	Negative modulator
Cabazitaxel	Small molecule	Inhibitor	Negative modulator
Cabozantinib	Small molecule	Inhibitor	Negative modulator
Caffeine	Small molecule	Antagonist	Negative modulator
Capecitabine	Small molecule	Inhibitor	Negative modulator
Carbamazepine	Small molecule	Blocker	Negative modulator
Carvedilol	Small molecule	Antagonist	Negative modulator
Cediranib	Small molecule	Inhibitor	Negative modulator
Celecoxib	Small molecule	Inhibitor	Negative modulator
Cenobamate	Small molecule	Inhibitor	Negative modulator
Cenobamate	Small molecule	Positive allosteric modulator	Positive modulator
Cetuximab	Antibody	Inhibitor	Negative modulator
Cholecalciferol	Small molecule	Agonist	Positive modulator
Cilengitide	Protein	Antagonist	Negative modulator
Cixutumumab	Antibody	Antagonist	Negative modulator
Clobazam	Small molecule	Positive allosteric modulator	Positive modulator
Clonazepam	Small molecule	Positive allosteric modulator	Positive modulator
Colchicine	Small molecule	Inhibitor	Negative modulator
Cudc-101	Small molecule	Inhibitor	Negative modulator
Cyclosporine	Protein	Modulator	Other
Cytarabine	Small molecule	Inhibitor	Negative modulator
Dactolisib	Small molecule	Inhibitor	Negative modulator
Dalfampridine	Small molecule	Blocker	Negative modulator
Dasatinib	Small molecule	Inhibitor	Negative modulator
Davunetide	Protein	Stabiliser	Other
Denileukin diftitox	Protein	Binding agent	Other
Denileukin diftitox	Protein	Inhibitor	Negative modulator
Desflurane	Small molecule	Opener	Positive modulator
Desflurane	Small molecule	Positive allosteric modulator	Positive modulator
Desflurane	Small molecule	Positive modulator	Positive modulator
Dexamethasone	Small molecule	Agonist	Positive modulator
Dexmedetomidine	Small molecule	Agonist	Positive modulator
Dextromethorphan	Small molecule	Agonist	Positive modulator
Dextromethorphan	Small molecule	Antagonist	Negative modulator
Diazepam	Small molecule	Positive allosteric modulator	Positive modulator
Diclofenac	Small molecule	Inhibitor	Negative modulator
Digoxin	Small molecule	Inhibitor	Negative modulator
Dipyridamole	Small molecule	Inhibitor	Negative modulator
Docetaxel	Small molecule	Inhibitor	Negative modulator

Dovitinib	Small molecule	Inhibitor	Negative modulator
Doxorubicin	Small molecule	Inhibitor	Negative modulator
Doxycycline	Small molecule	Inhibitor	Negative modulator
Dronedarone	Small molecule	Blocker	Negative modulator
Ds-7423	Small molecule	Inhibitor	Negative modulator
Emricasan	Small molecule	Inhibitor	Negative modulator
Enmd-981693	Small molecule	Inhibitor	Negative modulator
Entinostat	Small molecule	Inhibitor	Negative modulator
Epinephrine	Small molecule	Agonist	Positive modulator
Ergocalciferol	Small molecule	Agonist	Positive modulator
Eribulin	Small molecule	Inhibitor	Negative modulator
Erlosamide	Small molecule	Blocker	Negative modulator
Erlotinib	Small molecule	Inhibitor	Negative modulator
Eslicarbazepine acetate	Small molecule	Blocker	Negative modulator
Estrogens, conjugated	Small molecule	Agonist	Positive modulator
Etanercept	Protein	Inhibitor	Negative modulator
Etaracizumab	Antibody	Antagonist	Negative modulator
Etoposide	Small molecule	Inhibitor	Negative modulator
Everolimus	Small molecule	Inhibitor	Negative modulator
Famitinib	Small molecule	Inhibitor	Negative modulator
Filgrastim	Protein	Agonist	Positive modulator
Fingolimod	Small molecule	Agonist	Positive modulator
Fludarabine phosphate	Small molecule	Inhibitor	Negative modulator
Fluorouracil	Small molecule	Inhibitor	Negative modulator
Foretinib	Small molecule	Inhibitor	Negative modulator
Fosphenytoin	Small molecule	Blocker	Negative modulator
Fulvestrant	Small molecule	Antagonist	Negative modulator
Gaboxadol	Small molecule	Agonist	Positive modulator
Ganaxolone	Small molecule	Positive allosteric modulator	Positive modulator
Ganetespib	Small molecule	Inhibitor	Negative modulator
Gefitinib	Small molecule	Inhibitor	Negative modulator
Gemcitabine	Small molecule	Inhibitor	Negative modulator
Hydroxychloroquine	Small molecule	Antagonist	Negative modulator
Ibuprofen	Small molecule	Inhibitor	Negative modulator
Imatinib	Small molecule	Inhibitor	Negative modulator
Imiquimod	Small molecule	Agonist	Positive modulator
Indomethacin	Small molecule	Inhibitor	Negative modulator
Ipilimumab	Antibody	Inhibitor	Negative modulator
Irinotecan	Small molecule	Inhibitor	Negative modulator
Isotretinoin	Small molecule	Agonist	Positive modulator
Ixabepilone	Small molecule	Inhibitor	Negative modulator
Ketamine	Small molecule	Negative allosteric modulator	Negative modulator
Lamotrigine	Small molecule	Blocker	Negative modulator
Lapatinib	Small molecule	Inhibitor	Negative modulator

Lenalidomide	Small molecule	Inhibitor	Negative modulator
Lenvatinib	Small molecule	Inhibitor	Negative modulator
Lestaurtinib	Small molecule	Inhibitor	Negative modulator
Levetiracetam	Small molecule	Blocker	Negative modulator
Levetiracetam	Small molecule	Modulator	Other
Lidocaine	Small molecule	Blocker	Negative modulator
Linifanib	Small molecule	Inhibitor	Negative modulator
Lithium carbonate	Small molecule	Inhibitor	Negative modulator
Lorazepam	Small molecule	Positive allosteric modulator	Positive modulator
Marimastat	Small molecule	Inhibitor	Negative modulator
Masitinib	Small molecule	Inhibitor	Negative modulator
Me-344	Small molecule	Inhibitor	Negative modulator
Melatonin	Small molecule	Agonist	Positive modulator
Memantine	Small molecule	Negative allosteric modulator	Negative modulator
Mepivacaine	Small molecule	Blocker	Negative modulator
Metformin	Small molecule	Inhibitor	Negative modulator
Methotrexate	Small molecule	Inhibitor	Negative modulator
Methylprednisolone	Small molecule	Agonist	Positive modulator
Mexiletine	Small molecule	Blocker	Negative modulator
Midazolam	Small molecule	Positive allosteric modulator	Positive modulator
Midostaurin	Small molecule	Inhibitor	Negative modulator
Mirvetuximab soravtansine	Antibody	Binding agent	Other
Mirvetuximab soravtansine	Antibody	Inhibitor	Negative modulator
Mk-2206	Small molecule	Inhibitor	Negative modulator
Motesanib	Small molecule	Inhibitor	Negative modulator
Motexafin gadolinium	Small molecule	Inhibitor	Negative modulator
Muparfostat	Oligosaccharide	Inhibitor	Negative modulator
Mycophenolate mofetil	Small molecule	Inhibitor	Negative modulator
Naltrexone	Small molecule	Antagonist	Negative modulator
Nerispirdine	Small molecule	Blocker	Negative modulator
Nintedanib	Small molecule	Inhibitor	Negative modulator
Obatoclax	Small molecule	Inhibitor	Negative modulator
Octreotide	Protein	Agonist	Positive modulator
Olanzapine	Small molecule	Antagonist	Negative modulator
Olaparib	Small molecule	Inhibitor	Negative modulator
Omeprazole	Small molecule	Inhibitor	Negative modulator
Oxcarbazepine	Small molecule	Blocker	Negative modulator
Paclitaxel	Small molecule	Inhibitor	Negative modulator
Paclitaxel poliglumex	Unknown	Stabiliser	Other
Panitumumab	Antibody	Inhibitor	Negative modulator
Panobinostat	Small molecule	Inhibitor	Negative modulator
Pasireotide	Protein	Agonist	Positive modulator
Pazopanib	Small molecule	Inhibitor	Negative modulator

Pegfilgrastim	Protein	Agonist	Positive modulator
Pembrolizumab	Antibody	Inhibitor	Negative modulator
Pemetrexed	Small molecule	Inhibitor	Negative modulator
Pentoxifylline	Small molecule	Antagonist	Negative modulator
Pentoxifylline	Small molecule	Inhibitor	Negative modulator
Perampanel	Small molecule	Antagonist	Negative modulator
Pf-06372865	Small molecule	Agonist	Positive modulator
Phenytoin	Small molecule	Blocker	Negative modulator
Pictilisib	Small molecule	Inhibitor	Negative modulator
Pilaralisib	Small molecule	Inhibitor	Negative modulator
Pioglitazone	Small molecule	Agonist	Positive modulator
Pomalidomide	Small molecule	Inhibitor	Negative modulator
Prednisone	Small molecule	Agonist	Positive modulator
Procainamide	Small molecule	Blocker	Negative modulator
Propafenone	Small molecule	Antagonist	Negative modulator
Propafenone	Small molecule	Blocker	Negative modulator
Propofol	Small molecule	Positive allosteric modulator	Positive modulator
Propranolol	Small molecule	Antagonist	Negative modulator
Quinidine	Small molecule	Blocker	Negative modulator
Regorafenib	Small molecule	Inhibitor	Negative modulator
Riluzole	Small molecule	Blocker	Negative modulator
Risperidone	Small molecule	Antagonist	Negative modulator
Rituximab	Antibody	Binding agent	Other
Rocuronium	Small molecule	Antagonist	Negative modulator
Roflumilast	Small molecule	Inhibitor	Negative modulator
Romidepsin	Protein	Inhibitor	Negative modulator
Rontalizumab	Antibody	Inhibitor	Negative modulator
Ropivacaine	Small molecule	Blocker	Negative modulator
Rosiglitazone	Small molecule	Agonist	Positive modulator
Rosuvastatin	Small molecule	Inhibitor	Negative modulator
Rufinamide	Small molecule	Blocker	Negative modulator
Ruxolitinib	Small molecule	Inhibitor	Negative modulator
Sagopilone	Small molecule	Stabiliser	Other
Samotolisib	Small molecule	Inhibitor	Negative modulator
Sargramostim	Protein	Agonist	Positive modulator
Selumetinib	Small molecule	Inhibitor	Negative modulator
Selurampanel	Small molecule	Antagonist	Negative modulator
Semaxanib	Small molecule	Inhibitor	Negative modulator
Sevoflurane	Small molecule	Opener	Positive modulator
Sevoflurane	Small molecule	Positive allosteric modulator	Positive modulator
Sevoflurane	Small molecule	Positive modulator	Positive modulator
Sifalimumab	Antibody	Inhibitor	Negative modulator
Simvastatin	Small molecule	Inhibitor	Negative modulator
Sorafenib	Small molecule	Inhibitor	Negative modulator
Sotalol	Small molecule	Antagonist	Negative modulator
Sotalol	Small molecule	Blocker	Negative modulator

Sotrastaurin	Small molecule	Inhibitor	Negative modulator
Sunitinib	Small molecule	Inhibitor	Negative modulator
Tacedinaline	Small molecule	Inhibitor	Negative modulator
Tacrolimus	Small molecule	Inhibitor	Negative modulator
Tanespimycin	Small molecule	Inhibitor	Negative modulator
Taselisib	Small molecule	Inhibitor	Negative modulator
Tedisamil	Small molecule	Blocker	Negative modulator
Temsirolimus	Small molecule	Inhibitor	Negative modulator
Thalidomide	Small molecule	Inhibitor	Negative modulator
Theophylline	Small molecule	Antagonist	Negative modulator
Theophylline	Small molecule	Inhibitor	Negative modulator
Tipifarnib	Small molecule	Inhibitor	Negative modulator
Tivozanib	Small molecule	Inhibitor	Negative modulator
Tofacitinib	Small molecule	Inhibitor	Negative modulator
Topiramate	Small molecule	Antagonist	Negative modulator
Topiramate	Small molecule	Blocker	Negative modulator
Topiramate	Small molecule	Inhibitor	Negative modulator
Topiramate	Small molecule	Positive modulator	Positive modulator
Topotecan	Small molecule	Inhibitor	Negative modulator
Trastuzumab	Antibody	Inhibitor	Negative modulator
Trastuzumab emtansine	Antibody	Inhibitor	Negative modulator
Tretinoin	Small molecule	Agonist	Positive modulator
Ucn-01	Small molecule	Inhibitor	Negative modulator
Ustekinumab	Antibody	Inhibitor	Negative modulator
Valproic acid	Small molecule	Inhibitor	Negative modulator
Vandetanib	Small molecule	Inhibitor	Negative modulator
Vatalanib	Small molecule	Inhibitor	Negative modulator
Veliparib	Small molecule	Inhibitor	Negative modulator
Verapamil	Small molecule	Blocker	Negative modulator
Vernakalant	Small molecule	Blocker	Negative modulator
Vinblastine	Small molecule	Inhibitor	Negative modulator
Vincristine	Small molecule	Inhibitor	Negative modulator
Vinflunine	Small molecule	Inhibitor	Negative modulator
Vinorelbine	Small molecule	Inhibitor	Negative modulator
Vorinostat	Small molecule	Inhibitor	Negative modulator
Voxtalisib	Small molecule	Inhibitor	Negative modulator
Warfarin	Small molecule	Inhibitor	Negative modulator
Zonisamide	Small molecule	Blocker	Negative modulator

III- Drugs associated to Target Classes

Table that associates each drug of our data to own target class

Drug	Target class
Abatacept	· Surface antigen
Abiraterone	· Nuclear hormone receptor subfamily 3 group c member 4
Abiraterone	· Oxidoreductasecytochrome p450 17a1
Acamprosate	· Nmda receptor
Acamprosate	· Gaba-A receptor
Acetaminophen	· Hydrolaseoxidoreductase
Acetaminophen	· Transient receptor potential channel
Acitretin	· Nuclear hormone receptor subfamily 1 group b member 1
Afatinib	· Tyrosine protein kinase EGFR family
Aflibercept	· Secreted protein
Aldesleukin	· Membrane receptor
Alitretinoin	· Nuclear hormone receptor subfamily 1 group b member 1
Alprazolam	· Gaba-A receptor
Alvocidib	· CMGC PROTEIN KINASE CDC2 SUBFAMILY · CMGC PROTEIN KINASE CDK9 SUBFAMILY · CMGC PROTEIN KINASE GROUP · CMGC PROTEIN KINASE CDK7 SUBFAMILY
Amantadine	· NMDA RECEPTOR
Amifampridine	· Voltage-gated potassium channel
Amiselimod	· EDG receptor
Amisulpride	· 5HT3 receptordopamine receptor
Amlodipine	· Voltage-gated calcium channel
Anlotinib	· Tyrosine protein kinase VEGFR family · Tyrosine protein kinase PDGFR family
Apitolisib	· Enzyme atypical protein kinase FRAP subfamily
Apremilast	· Phosphodiesterase 4A
5Aripiprazole	· Serotonin receptor
Aripiprazole	· Dopamine receptorserotonin receptor
Asm-024	· Acetylcholine receptor · Nicotinic acetylcholine receptor alpha subunit
Asm-024	· Nicotinic acetylcholine receptor alpha subunit
Aspirin	· Oxidoreductase
Atorvastatin	· Oxidoreductase
Axitinib	· Tyrosine protein kinase VEGFR family
Azacitidine	· DNA methyltransferase
Azd-4547	· Tyrosine protein kinase FGFR family
Azd1305	· Voltage-gated calcium channel · Voltage-gated sodium channel · Voltage-gated potassium channel
Becatecarin	· Isomerase
Belinostat	· HDAC class i
Bevacizumab	· Secreted protein
Bexarotene	· Nuclear hormone receptor subfamily 2 group b member 1

Binimetinib	· STE protein kinase STE7 family
Birabresib	· Bromodomain
Bms-690514	· Tyrosine protein kinase VEGFR family · Tyrosine protein kinase EGFR family
Bortezomib	· Threonine protease t1A subfamily
Brentuximab vedotin	· Structural proteinsecreted protein
Brexanolone	· Gaba-A receptor
Brivanib	· Tyrosine protein kinase VEGFR family · Tyrosine protein kinase FGFR family
Buparlisib	· Enzyme
Cabazitaxel	· Structural protein
Cabozantinib	· Tyrosine protein kinase MET family · Tyrosine protein kinase VEGFR family
Caffeine	· Adenosine receptor
Capecitabine	· Transferase
Carbamazepine	· Voltage-gated sodium channel
Carvedilol	· Adrenergic receptor
Cediranib	· Tyrosine protein kinase VEGFR family · Tyrosine protein kinase PDGFR family
Celecoxib	· Oxidoreductase
Cenobamate	· Voltage-gated sodium channel
Cenobamate	· Gaba-A receptor
Cetuximab	· Tyrosine protein kinase EGFR family
Cholecalciferol	· Nuclear hormone receptor subfamily 1 group i member 1
Cilengitide	· Membrane receptor
Cixutumumab	· Tyrosine protein kinase INSR family
Clobazam	· Gaba-A receptor
Clonazepam	· Gaba-A receptor
Colchicine	· Structural protein
Cudc-101	· HDAC class ityrosine protein kinase EGFR family
Cyclosporine	· Isomerase
Cytarabine	· Enzyme
Dactolisib	· Enzymeatypical protein · Kinase FRAP subfamily
Dalfampridine	· Voltage-gated potassium channel
Dasatinib	· Tyrosine protein kinase PDGFR family · Tyrosine protein kinase SRC family · Tyrosine protein kinase ABL family · Tyrosine protein kinase EPH family
Davunetide	· Structural protein
Denileukin diftitox	· Membrane receptor
Denileukin diftitox	· Other cytosolic protein
Desflurane	· Two-pore domain potassium channel
Desflurane	· Gaba-A receptor
Desflurane	· Glycine receptor
Dexamethasone	· Nuclear hormone receptor subfamily 3 group c member 1

Dexmedetomidine	· Adrenergic receptor
Dextromethorphan	· Membrane receptor
Dextromethorphan	· NMDA receptor
Diazepam	· Gaba-A receptor
Diclofenac	· Oxidoreductase
Digoxin	· Hydrolase
Dipyridamole	· Phosphodiesterase 10aslc29 facilitative nucleoside transporter family
Docetaxel	· Structural protein
Dovitinib	· Tyrosine protein kinase FGFR family · Tyrosine protein kinase PDGFR family · Tyrosine protein kinase VEGFR family
Doxorubicin	· Isomerase
Doxycycline	· Metallo protease M10a subfamily
Dronedarone	· Voltage-gated sodium channel · Inwardly rectifying potassium channel · Cyclic nucleotide-regulated channel · Voltage-gated calcium channel
Ds-7423	· Enzymeatypical protein kinase frap subfamily
Emricasan	· Cysteine protease c14 family
Enmd-981693	· Tyrosine protein kinase src family · Tyrosine protein kinase pdgfr family · other protein kinase AUR family · Tyrosine protein kinase FGFR family · Tyrosine protein kinase VEGFR family
Entinostat	· HDAC class i
Epinephrine	· Adrenergic receptor
Ergocalciferol	· Nuclear hormone receptor subfamily 1 group i member 1
Eribulin	· Structural protein
Erlosamide	· Voltage-gated sodium channel
Erlotinib	· Tyrosine protein kinase EGFR family
Eslicarbazepine acetate	· Voltage-gated sodium channel
Estrogens, conjugated	· Nuclear hormone receptor subfamily 3 group a member 1
Etanercept	· Secreted protein
Etaracizumab	· Membrane receptor
Etoposide	· Isomerase
Everolimus	· Isomerase
Famitinib	· Tyrosine protein kinase PDGFR family · Tyrosine protein kinase VEGFR family
Filgrastim	· Membrane receptor
Fingolimod	· Edg receptor
Fludarabine phosphate	· Enzymeoxidoreductase
Fluorouracil	· Transferase

Foretinib	· Tyrosine protein kinase PDGFR family · Tyrosine protein kinase TIE family · Tyrosine protein kinase VEGFR family · Tyrosine protein kinase MET family
Fosphenytoin	· Voltage-gated sodium channel
Fulvestrant	· Nuclear hormone receptor subfamily 3 group a member 1
Gaboxadol	· Gaba-A receptor
Ganaxolone	· Gaba-A receptor
Ganetespib	· Other cytosolic protein
Gefitinib	· Tyrosine protein kinase EGFR family
Gemcitabine	· Enzyme oxidoreductase
Hydroxychloroquine	· Toll-like and il-1 receptors
Ibuprofen	· Oxidoreductase
Imatinib	· Tyrosine protein kinase ABL family · Tyrosine protein kinase PDGFR family
Imiquimod	· Toll-like and il-1 receptors
Indomethacin	· Oxidoreductase
Ipilimumab	· Surface antigen
Irinotecan	· Isomerase
Isotretinoin	· Nuclear hormone receptor subfamily 1 group b member 1
Ixabepilone	· Structural protein
Ketamine	· NMDA receptor
Lamotrigine	· Voltage-gated sodium channel
Lapatinib	· Tyrosine protein kinase EGFR family
Lenalidomide	· Unclassified protein
Lenvatinib	· Tyrosine protein kinase VEGFR family
Lestaurtinib	· Tyrosine protein kinase PDGFR family · Tyrosine protein kinase RET family · Tyrosine protein kinase TRK family · Tyrosine protein kinase JAKA family
Levetiracetam	· Voltage-gated calcium channel
Levetiracetam	· Vesicular neurotransmitter transporter family
Lidocaine	· Voltage-gated sodium channel
Linifanib	· Tyrosine protein kinase PDGFR family · Tyrosine protein kinase VEGFR family
Lithium carbonate	· CMGC protein kinase GSK family · Hydrolase
Lorazepam	· Gaba-A receptor
Marimastat	· Metallo protease m10a subfamily
Masitinib	· Tyrosine protein kinase PDGFR family · Tyrosine protein kinase FGFR family
Me-344	· Oxidoreductase
Melatonin	· Melatonin receptor
Memantine	· NMDA receptor
Mepivacaine	· Voltage-gated sodium channel
Metformin	· Oxidoreductase
Methotrexate	· Oxidoreductase

Methylprednisolone	· Nuclear hormone receptor subfamily 3 group c member 1
Mexiletine	· Voltage-gated sodium channel
Midazolam	· Gaba-A receptor
Midostaurin	· AGC protein kinase PKC alpha subfamily · Tyrosine protein kinase PDGFR family · Tyrosine protein kinase VEGFR family
Mirvetuximab soravtansine	· Membrane receptor
Mirvetuximab soravtansine	· Structural protein
Mk-2206	· AGC protein kinase AKT family
Motesanib	· Tyrosine protein kinase PDGFR family · Tyrosine protein kinase RET family · Tyrosine protein kinase VEGFR family
Motexafin gadolinium	· Enzyme oxidoreductase
Muparfostat	· Hydrolasesecreted protein
Mycophenolate mofetil	· Oxidoreductase
Naltrexone	· Opioid receptor
Nerispirdine	· Voltage-gated sodium channel · Voltage-gated potassium channel
Nintedanib	· Tyrosine protein kinase VEGFR family · Tyrosine protein kinase FGFR family · Tyrosine protein kinase PDGFR family
Obatoclax	· Bcl-2 family · Cytosolic protein
Octreotide	· Somatostatin receptor
Olanzapine	· Serotonin receptor · Dopamine receptor
Olaparib	· Transferase
Omeprazole	· Hydrogen potassium ATPase
Oxcarbazepine	· Voltage-gated sodium channel
Paclitaxel	· Structural protein
Paclitaxel poliglumex	· Structural protein
Panitumumab	· Tyrosine protein kinase EGFR family
Panobinostat	· HDAC class I
Pasireotide	· Somatostatin receptor
Pazopanib	· Tyrosine protein kinase TEC family · Tyrosine protein kinase VEGFR family · Tyrosine protein kinase PDGFR family · Tyrosine protein kinase SRC family · Tyrosine protein kinase FGFR family
Pegfilgrastim	· Membrane receptor
Pembrolizumab	· Surface antigen
Pemetrexed	· Oxidoreductase · Ligase · Transferase
Pentoxifylline	· Adenosine receptor
Pentoxifylline	· Phosphodiesterase 10a
Perampanel	· AMPA receptor

Pf-06372865	· Gaba-A receptor
Phenytoin	· Voltage-gated sodium channel
Pictilisib	· Enzyme
Pilaralisib	· Enzyme
Pioglitazone	· Nuclear hormone receptor subfamily 1 group c member 3
Pomalidomide	· Unclassified protein
Prednisone	· Nuclear hormone receptor subfamily 3 group c member 1
Procainamide	· Voltage-gated sodium channel
Propafenone	· Adrenergic receptor
Propafenone	· Voltage-gated sodium channel
Propofol	· Gaba-A receptor
Propranolol	· Adrenergic receptor
Quinidine	· Voltage-gated sodium channel
Regorafenib	· Tyrosine protein kinase fgfr family · Tyrosine protein kinase ddr family · Tyrosine protein kinase vegfr family TKL protein kinase RAF family · Tyrosine protein kinase ABL family · CMGC protein kinase P38 subfamily · Tyrosine protein kinase PDGFR family · Tyrosine protein kinase TIE family · Tyrosine protein kinase RET family · Tyrosine protein kinase SRC family
Riluzole	· Voltage-gated sodium channel
Risperidone	· Dopamine receptor · Serotonin receptor
Rituximab	· Cd20 ca2+ channel family
Rocuronium	· Nicotinic acetylcholine receptor alpha subunit
Roflumilast	· Phosphodiesterase 4a
Romidepsin	· HDAC class i
Rontalizumab	· Secreted protein
Ropivacaine	· Voltage-gated sodium channel
Rosiglitazone	· Nuclear hormone receptor subfamily 1 group c member 3
Rosuvastatin	· Oxidoreductase
Rufinamide	· Voltage-gated sodium channel
Ruxolitinib	· Tyrosine protein kinase JAKA family
Sagopilone	· Structural protein
Samotolisib	· Enzymeatypical protein kinase FRAP subfamily
Sargramostim	· Membrane receptor
Selumetinib	· STE protein kinase STE7 family
Selurampanel	· Kainate receptor · AMPA receptor
Semaxanib	· Tyrosine protein kinase PDGFR family · Tyrosine protein kinase VEGFR family
Sevoflurane	· Two-pore domain potassium channel
Sevoflurane	· Gaba-a receptor
Sevoflurane	· Glycine receptor

Sifalimumab	· Secreted protein
Simvastatin	· Oxidoreductase
Sorafenib	· TKL protein kinase RAF family · Tyrosine protein kinase PDGFR family · Tyrosine protein kinase VEGFR family · Tyrosine protein kinase RET family
Sotalol	· Adrenergic receptor
Sotalol	· Voltage-gated potassium channel
Sotrastaurin	· AGC protein kinase PKC alpha subfamily
Sunitinib	· Tyrosine protein kinase PDGFR family · Tyrosine protein kinase VEGFR family · Tyrosine protein kinase RET family
Tacedinaline	· HDAC class i
Tacrolimus	· Isomerase
Tanespimycin	· Other cytosolic protein
Taselisib	· Enzyme
Tedisamil	· Voltage-gated potassium channel
Temsirolimus	· Isomerase
Thalidomide	· Unclassified protein
Theophylline	· Adenosine receptor
Theophylline	· Phosphodiesterase 4aphosphodiesterase 3a
Tipifarnib	· Transferase
Tivozanib	· Tyrosine protein kinase VEGFR family
Tofacitinib	· Tyrosine protein kinase JAKA family
Topiramate	· AMPA receptorkainate receptor
Topiramate	· Voltage-gated sodium channel
Topiramate	· Lyase
Topiramate	· Gaba-A receptor
Topotecan	· Isomerase
Trastuzumab	· Tyrosine protein kinase EGFR family
Trastuzumab emtansine	· Structural proteintyrosine protein kinase EGFR family
Tretinoin	· Nuclear hormone receptor subfamily 1 group b member 1
Ucn-01	· AGC protein kinase PKC alpha subfamily · CMGC protein kinase CDC2 subfamily · CMGC protein kinase group · AGC protein kinase PDK1 subfamily · CAMK protein kinase CHK1 subfamily
Ustekinumab	· Secreted protein
Valproic acid	· Oxidoreductase
Vandetanib	· Kinasetyrosine protein kinase EGFR family · Tyrosine protein kinase VEGFR family · Tyrosine protein kinase SRC family · Tyrosine protein kinase RET family · Tyrosine protein kinase TIE family
Vatalanib	· Tyrosine protein kinase PDGFR family · Tyrosine protein kinase VEGFR family
Veliparib	· Transferase
Verapamil	· Voltage-gated calcium channel

Vernakalant	· Voltage-gated potassium channel · Inwardly rectifying potassium channel · Voltage-gated sodium channel
Vinblastine	· Structural protein
Vincristine	· Structural protein
Vinflunine	· Structural protein
Vinorelbine	· Structural protein
Vorinostat	· HDAC class ihdac class IIb
Voxtalisib	· Enzyme atypical protein kinase FRAP subfamily
Warfarin	· Oxidoreductase
Zonisamide	· Voltage-gated sodium channel