T.C.
FATIH UNIVERSITY
INSTITUTE OF BIOMEDICAL ENGINEERING

DATA MINING APPROACHES TO DRUG REPOSITIONING TO MULTIPLE DISEASES

ABDULLAH ALRHMOUN

MSc THESIS
BIOMEDICAL ENGINEERING PROGRAMME

THESIS ADVISOR
ASST. PROF. DR. AYDIN ALBAYRAK

İSTANBUL, JANUARY / 2016

T.C.
FATIH UNIVERSITY
INSTITUTE OF BIOMEDICAL ENGINEERING

DATA MINING APPROACHES TO DRUG REPOSITIONING TO MULTIPLE DISEASES

ABDULLAH ALRHMOUN

MSc THESIS
BIOMEDICAL ENGINEERING PROGRAMME

THESIS ADVISOR
ASST. PROF. DR. AYDIN ALBAYRAK

İSTANBUL, JANUARY / 2016

T.C.
FATİH ÜNİVESİTESİ
BİYOMEDİKAL MÜHENDİSLİK ENSTİTÜSÜ

VERİ MADENCİLİĞİ YÖTEMLERİ KULLANILARAK VAR OLAN İLAÇLARIN FARKLI HASTALIKLAR İÇİN YENİDEN TASARLANMASI

ABDULLAH ALRHMOUN

YÜKSEK LİSANS TEZİ
BİYOMEDİKAL MÜHENDİSLİĞİ PROGRAMI

DANIŞMAN
YRD. DOÇ. DR. AYDIN ALBAYRAK

İSTANBUL, OCAK / 2016

T.C.

FATIH UNIVERSITY

INSTITUTE OF BIOMEDICAL ENGINEERING

Abdullah Alrhmoun, a **MSc** student of Fatih University **Institute of Biomedical Engineering** student ID **520113004**, successfully defended the **thesis/dissertation** entitled "**Data mining approaches to drug repositioning to multiple diseases**", which he prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Committee Members

Thesis Advisor: **Asst. Prof. Dr. Aydın ALBAYRAK**
Fatih University

Jury Members: **Prof. Dr. Osman Uğur SEZERMAN**
Acıbadem University

Asst. Prof. Dr. Haşim Özgür TABAKOĞLU
Fatih University

It is approved that this thesis has been written in compliance with the formatting rules laid down by the Institute of Biomedical Engineering.

Prof. Dr. Sadık KARA
Director

Date of Submission: 13 January 2016
Date of Defense: 25 January 2016

To my first teacher: my mother,

This study was supported by Fatih University Research and Development Management office

ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my advisor Dr. Aydın Albayrak for his continuous support during my study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me through the time of research and writing of this thesis.

Additionally, I would like to extend my gratitude to my family for their endless support, love, and especially for their patience during my thesis.

February 2016 ABDULLAH ALRHMOUN

TABLE OF CONTENTS

Page

LIST OF SYMBOLS

~	Approximation tilde
f	Formula
$\cup$	Set-theoretic union
$\cap$	Set-theoretic intersection
$\otimes$	Tensor product
Φ	Phi
$\in$	Set membership

ABBREVIATIONS

ADR : Adverse Drug Reaction
CMap : Connectivity Map
CPI : Chemical Protein Interactome
DR : Drug Repositioning
FDA : Food and Drug Administration
GA : Genetic Algorithm
GO : Gene Ontology
IC : Incremental Construction
KEGG : Kyoto Encyclopedia of Genes and Genomes
MA : Matching Algorithms
MC : Monte Carlo
MCSS : Multiple Copy Simultaneous Search
MD : Molecular Dynamics
ML : Machine Learning
MoA : Mode of Action
PDB : Protein Data Bank
NCBI : National Center for Biotechnology Information
NLP : Natural Language Processing
PCA : Principal Component Analysis
PPI : Protein-Protein Interactions
QSAR : Quantitative Structure-Activity Relationship
RFE : Recursive Feature Elimination
SNS : Shared Neighborhood Scoring
SVM : Support Vector Machine
UMLS : Unified Medical Language System

LIST OF FIGURES

Page

LIST OF TABLES

Page

SUMMARY

DATA MINING APPROACHES TO DRUG REPOSITIONING TO MULTIPLE DISEASES

Abdullah ALRHMOUN

Biomedical Engineering Programme

MSc Thesis

Advisor: Assist. Prof. Dr. Aydın ALBAYRAK

Drug repositioning is defined as the identification of new uses for existing drugs. The ultimate goal is to reduce time and costs associated with the traditional drug development process. In recent years, drug repositioning has garnered the attention of both pharmaceutical companies and academic research centers.

In this study, drug and disease related data such as substructures, side effects, target protein and miRNA from a variety of online databases have been collected and compiled into feature matrix with 639 known drug-disease associations and 1647 drug-disease related features. R language was used for cleaning and preparing the compiled data for analysis whereas numerous Python packages were used for applying the SVM classification routine to select features with better predictive potentials in drug-repositioning. A classification accuracy of 99% has been achieved for drug repositioning with as few as 20 features which contain a conserved subgroup of chemical substructures and miRNAs.

Keywords: *Drug repositioning, Data mining, Machine learning, classification, support vector machine*

FATIH UNIVERSITY – INSTITUTE OF BIOMEDICAL ENGINEERING

ÖZET

VERİ MADENCİLİĞİ YÖTEMLERİ KULLANILARAK VAR OLAN İLAÇLARIN FARKLI HASTALIKLAR İÇİN YENİDEN TASARLANMASI

Abdullah ALRHMOUN

Biyomedikal Mühendisliği Programı

Yüksek Lisans Tezi

Danışman: Yrd. Doç. Dr. Aydın ALBAYRAK

İlaç repozisyonu var olan bir ilacın yeni kullanım alanlarını bulma süreci olarak tanımlanır. Bu işlemdeki asıl amaç geleneksel ilaç geliştirme süresini ve maliyetini azaltmaktır. Özellikle son yıllarda ilaç repozisyonu ilaç şirketlerinin ve akademik araştırma grupların yoğun ilgisini çekmiştir.

Bu çalışmada ilaç-hastalık ilişkisini tanımlama da kimyasal altyapıları, yan etkiler, hedef protein ve miRNA gibi veriler değişik veribankalarından derlenerek 639 ilaç-hastalık etkileşimi ve 1647 ilaç veya hastalık ilişkisine dair özellik içeren bir matriks oluşturulmuştur. İlaç repozisyon potansiyeli en yüksek olan ilaçların belirlenmesi işlemi sırasında kullanılan verilerin analize hazır hale getirilmesi için R yazım dili ile *Destek Vektör Makineleri* (SVM) yöntemi ile sınıflandırma işlemi sırasında birçok Python programcıkları kullanılmıştır. İlk defa kimyasal altyapılar ve miRNA gibi sadece 20 adet özellik kullanılarak sınıflandırma işlemi sırasında %99 doğruluk oranı elde edilmiştir.

Anahtar kelimeler: *İlaç repozisyonu, Veri madenciliği, Makine öğrenimi, Sınıflandırma, Destek Vektör Makinesi*

FATİH ÜNİVERSİTESİ –BİYOMEDİKAL MÜHENDİSLİK ENSTİTÜSÜ

CHAPTER 1

INTRODUCTION

1.1 Purpose of Thesis

‘The most fruitful basis for the discovery of a new drug is to start with an old drug’ [57]

-Noble laureate James Black-

Drug repositioning is a collection of important strategic steps in drug development and discovery. It has a great potential to push the boundaries in drug related scientific researches and pharmaceutical industry companies’ revenues by reducing time and costs associated with new drug development. Works on drug repositioning have been achieved initially by noticing unexpected results that happen when experimenting a drug during development periods or trial stages. Serendipity was the most important factor during the early drug repositioning works and produced successful repositioned drugs like Sildenafil.

Besides the experimental research and noticed signs, the information revolution which dominates this century helped to increase the understanding of biological systems and provide the researchers with fruitful and useful data. Nevertheless, biological systems are highly complex and fuzzy and still cannot be completely understood.

Interestingly, drug repositioning in the context of bioinformatics has a big potential to benefit from the rising number of data generated and provided through biological, chemical, biophysical, and genomic studies, or via the interactions between any of these disciplines. There is seemingly a proportional relationship between data and technology, where the rise in data produced is reflected on the development of more accurate and precise technologies to help understand the generated data. Hence, improved technology can help produce more clean and useful data.

Network analysis, machine learning, and natural language processing are the most successful techniques to extract information from digital data, and transform it into information. All these techniques have been used on drug repositioning research and generated important results, such as nominating new drug candidates and speeding up the drug development process.

Computational drug repositioning which depends essentially on available data, benefits from open source and public databases. Such databases can provide either general information about various aspects and types of data or specific specialized databases on a single type of data. Examples of specialized databases are: side effect data available on SIDER [94], chemical structure data available on PubChem [95], and genes and genetic disorders available on OMIM [96].

The main objectives of the thesis can be summarized in the following points:

1. Repositioning existing and selected drugs into different diseases, through drug-target, multiple targets, or other properties and associations.

2. Enriching the drug repositioning and discovery research area, and to be part of pharmaceutical bioinformatics non-profit academic researches.

3. Generate repositionable drug candidates that can be later experimentally researched for clinical use.

4. Apply machine learning algorithms to drug related bioinformatics research.

5. Bringing the attention of the bioinformatics community in Turkey into new, fruitful, applicable, and producible area of search.

In this thesis, I am going to use the most proper tools of data mining and machine learning to analyze online available data sets related to drug repositioning. It is an effort expended to facilitate new drug repositioning candidates, pushing the scientific research one step forward.

1.2 **Thesis Overview**

This thesis is organized into four chapters in addition to the references. **Chapter 1** is a general introduction to the topic, its importance, benefits, and the purpose of the thesis. **Chapter 2** defines the topic, previous examples, its concepts and methods, and general literature review. **Chapter 3** explains the methods that the thesis based on, and which

data sets and databases have been used. **Chapter 4** shows the results, make a conclusion on the results.

CHAPTER 2

DRUG REPOSITIONING

2.1 **Definition**

Drug repositioning by definition is the process of identifying new uses for existing drugs [1]. Alternative names are commonly used for the same expression including: *drug repurposing*, *drug re-tasking*, *drug reprofiling*, *indication expansion*, and *therapeutic switching* [1].

These drugs can be:

1. Approved drugs used for usual medical indications.
2. Compounds or molecules which did not pass the clinical trails.
3. Projects have been stoped for many reasons.

Barrat and Frail [2] in 2013, revolutionized the definition of drug repositioning as "*renewing failed drugs and expanding successful ones"*.

It is a challenging task to find new modes to cure diseases, and lately it became a fundamental quesition in biomedical research. In the last year, with the raising importance of big data and the tools relating to it, bioinformatic approaches became required and promising to accurately predict drug targets for a disease.

Enormous data sets describing drug effects and new exploited targets have been published, resulting in a massive amount of information and large-scale molecular data publically available on-line in libraries of biomedical publications such as PubMed [95] (Figure 2.1).

Drug repositioning strategies can make use of a variety of data sources and data mining approaches. Drug repositioning can achieve successful results because of the improved technologies that can enable the analysis of large experimental data to find novel patterns or associations. Data mining methodologies have been widely used to extract

the knowledge from genomic, metabolic, chemical and proteomic data.

Depending on the outcome of the data analysis and the extracted information, novel and unknown relationships can be discovered. These discoveries can lead to broader information enrichment in areas such as target selection and potential drug repositioning.

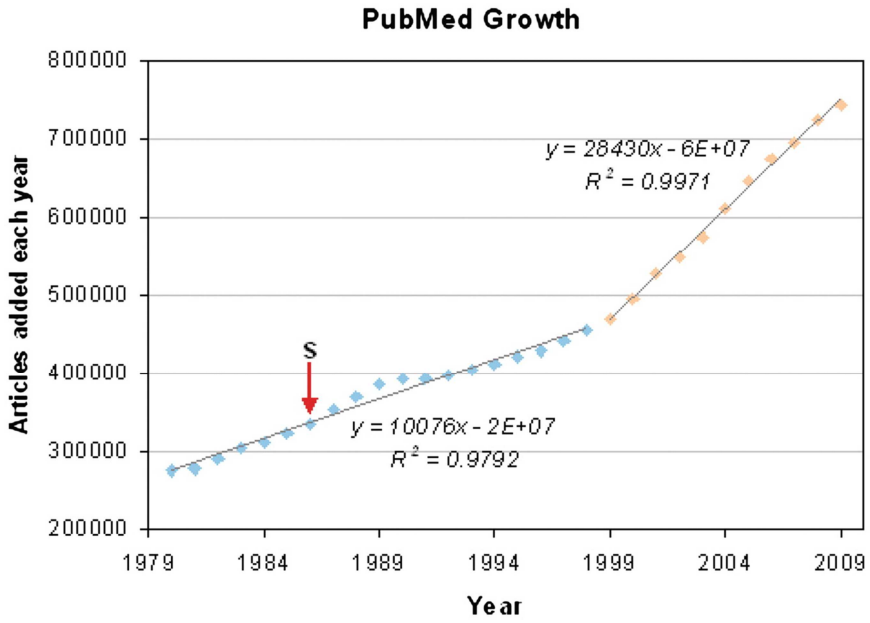


Figure 2.1 The number of articles added to PubMed each year is plotted against the year they were added. The growth rate increased almost 3-folds in the past 10 years [3]

The two main principles that usually rule the drug repositioning process are: i) the ability of one drug to affect several targets and ii) a disease specific target might be linked to another diseases or pathways. [4].

Based on these principles a shared gene or similar structures between two diseases, two drugs or a disease and a drug might be harnessed through some computational programs as candidates for drug repurposing [5].

The last decade witnessed a transition of experimental drug discovery from big pharmaceutical companies to startups, nonprofit organizations, and academic institutions [6], with special focus on rare diseases [7]. In the academic area, for example, there are several published studies from different universities [8], using a variety of computational and empirical methods, with high rate of positive results.

This transition was probably due to several factors: (1) the lack of new inventions

noticed in pharmaceutical sectors; (2) loosing a large number of skilled staff, due to the global economic deterioration which caused them to move to other research centers such as the academic institutes; (3) the establishment of three big initiatives in the US and Europe: a) Molecular Libraries Program [97], which supports research in chemical probe development projects. B) Clinical and Translational Science Award [98], which supports clinical and translational projects. C) Innovative Medicines Initiative [99], which combines pharmaceutical units and academic centers; (4) the last important factor is the rising amount of open source big data, tools, software, that supports drug discovery research projects [8].

Drug repositioning researchers have utilized several powerful methodologies such as: systems biology, network medicine, and bioinformatics approaches in an effort to determine unknown indications for existing drugs [9]. Nevertheless, until now, occasional observations of unexpected side effects of drugs under experimental testing or in the market were responsible for most of the successfully repositioned drugs.

Incorporation of advanced bioinformatics tools in drug repositioning studies, especially in the analysis of biomedical big data sets, led to a high quality and faster outcome. Analyzed data sets may consist of: gene expression profiles, chemical structure similarities, disease-drug network, literature mining, side-effect similarity, phenotypic disease network, disease comorbidity, pathway-based disease network and so on [9].

2.2 **Comparing Between Drug Repositioning and Drug Development**

A general comparison of novel drug development and drug repositioning based on several aspects and characteristics was provided in Table 2.1. This comparison serves the purpose of evaluating the advantages and disadvantages of the drug discovery research by both methods; whether in pharmaceutical companies or in non-profit research centers.

Table 2.1 A comparison between novel drug development and drug repositioning

Aspect	**Drug Development**	**Drug Repositioning**
Time	15 – 20 year per drug	3 – 10 year
Economic	Cost: 1$- 2$ billion per drug Fixed revenue	Reduce cost Extra revenue
Risk	Most candidate failed	Higher success rate

Scientific	Just 20 – 30 compounds each year	200 promising compounds

2.2.1 Time Aspect: *De novo* drug development is laborious; requires 15 – 20 years to bring a new chemical entity to market [10], while drug repositioning needs 3 – 10 years (60% lower) from indication identification to market [11]. Table 2.2 contains a summary of both drug repositioning and novel drug development pipelines from drug discovery and compound identification to approval and registration.

Table 2.2 Timeline of *de novo* drug development and drug repositioning

De novo drug development: Timeline to market					
Drug Discovery	Discovery & screening	Lead optimization	ADMET	Development	Registration
o Expression analysis o *In vitro* function o *In vitro* validation; for example: knockouts o Bioinformatics	Discovery: o Traditional o Combinatorial chemistry o Structure-based drug design Screening: o *In vitro* o *Ex vivo* and *in Vivo* o High throughput	o Traditional medicinal chemistry o Rational drug design	o Bioavailability systemic exposure (absorption, clearance and distribution)	o Must start clinical testing at phase 1 (phase 1/11 for cancer)	o US (FDA) o Europe (EMEA) o Japan (MHLW) o Rest of the world

Drug repositioning: Timeline to market			
Compound identification	Compound acquisition	Development	Registration
o Traditional medicinal chemistry o Rational drug design	o Licensing o Novel IP o Both licensing and novel IP o Internal sources	o May start at preclinical, phase 1 or phase 11 stages o Ability to leverage existing data packages	o US (FDA) o Europe (EMEA) o Japan (MHLW) o Rest of the world

2.2.2 Economic Aspect: The two important elements which suggest that drug repositioning is more advantageous economically over the *de novo* drug development

are: cost and revenue.

A. **Cost:** Developing a single drug requires 1$ - 2$ billion dollars [11] whilst repositioning an already existing drug costs much less because of reduced time and number of steps required for development.

B. **Revenue:** While drug development focuses on one target, drug repositioning takes in consideration several targets for a single drug. This will generate an extra revenue which can exceed billions. For instance, sales of sildenafil (brand name: viagra), repositioned for erectile dysfunction, reached US $1.88 billion annually, and thelidomide, repositioned drug for multiple myeloma and leprosy, had sales US $271 million in 2003 alone [11].

2.2.3 Risk Aspect: The success rate for developing new drug candidates is less than 10% [12], only 20 – 30 new chemical entities are approved per year in the US [13], and the development productivity has significantly declined in recent year [14]. On the other hand, repositioned drugs in the last few years account for about 30% of the new drugs that are approved and marketed. In addition to that, there are at least 200 compounds which are promising to be repositioned to fit many targets and diseases [15].

2.2.4 Scientific Aspect: Risk, economic, and time aspects are all scientific, but scientific here refers to the positive effect of drug repositioning on the scientific research. Drug repositioning research is usually carried by non-profitable instituations, and results in scientific knowledge expansion and positive medical outcomes. On the other hand, 90% of novel drug research and development are carried by profitable entities with commercial rather than scientific goals.

Rare diseases or the so called orphan diseases, are a group of diseases which affect a very small percentage of people. These diseases do not have usually persisting treatments, and because of its rareness no high cost drug development studies are usually performed. Nevertheless, many drug repositioning researchers consider these diseases in their research. In fact, the US food and drug administration (FDA) launched a database which incorporates all the reported disease and drug data about orphan diseases to facilitate drug repositioning studies for these diseases [16].

In summary, novel drug development is a time-consuming, expensive, and risky venture that requires coordinated multidisciplinary research in multiple stages with each requiring intensive and specialized resources.

2.3 Examples of Repositioned Drugs

It is important to understand the concepts and factors behind the successful findings in order to establish solid predictive models to repositioning of drugs. I will briefly talk about some of the successfully repositioned drugs, recent discoveries, and ongoing projects.

2.3.1 Most Successful Examples

1. Sildenafil: Also known as Viagra®, Sildenafil was developed in the late 1980s for the treatment of angina, the effect on angina was mild compared to the penile erections reported by most patients during the clinical trails as a side effect. The scientists decided to investigate the drug for this new indication by trying it on 3,700 men [17]. After their observation on the efficacy of the compound and the pharmacokinetic eligibility, the drug was repositioned for the treatment of erectile dysfunction. Viagra which started as a drug for angina then for erectile dysfunction was also confirmed for the treatment of pulmonary hypertension, making $1.88 billion each year of sales [18].

2. Raloxifene (brand name Evista): Initially, this drug was developed and studied initially to be used against breast cancer [11], but during the experimental studies the drug showed anti-oestrogenic effects [19]. In order to expand the production line and for commercial and strategic reasons, raloxifene was confirmed in 1999 as a unique indication for osteporosis. Then, in 2007 the drug was suggested and approved as a breast cancer preventive agent [20].

3. Thalidomide: The drug was developed and marketed to treat nausea in pregnancy. However, due to its disastrous side-effects which caused severe skeletal birth defects in over 15,000 infants, the drug was stoped [21]. Later on, thalidomide came back to market as the only drug approved for the treatment of rethema nodosum leprosum and multiple myeloma [22]. Now, the drug sales reach $271 million each year.

2.3.2 Recent Examples and Ongoing Projects

In Table 2.3, a list of drug repositioning examples and projects that show the original and novel indications for each drug has been provided.

Table 2.3 List of example projects in drug repositioning

Drug name	Original target	Original indication	New target	New indication	Reference
Duloxetine	Serotonin and norepinephrine reuptake	Depression	Serotonin and norepinephrine reuptake	Stress urinary incontinence, fibromyalgia chronic musculoskeletal pain	23
Imatinib	BCR-ABL	CML	KIT, PDGFRA	GIST	24
Raltegravir	-	HIV-1 integrase	-	Metnase; adjuvant therapy in cancer	8
Astemizole	-	Histamine HI receptors; Antihistamine for treatment of seasonal allergy	-	Inducer of autophagy; as adjuvant therapy in prostate cancer	8
Celecoxib	Cyclo oxygenase-2	-	Carbonic anhydrase	Glaucoma, cancer	25
Nelfinavir	HIV-1 protease	AIDS	Inhibits AKT pathway	Multiple disease	26
Minoxidil	Unknown	Hypertension	Unknown	Hair loss	27
Sunitinib	Multiple kinases	GIST, renal cell carcinoma	Unchanged	Pancreatic neuroendocrine tumors	28
Everolimus	mTOR	Immunosuppressant	Unchanged	Pancreatic neuroendocrine tumors	29
Phenothiazines	-	Prototype for neuroleptic drugs; antipsychotics for the management of schizophrenia	-	Anti-adhesion inhibitors against inflammation and cancer	8
Trastuzumab	HER2	HER2-positive breast cancer	Unchanged	HER2-positive metastatic gastric cancer	30

2.4 General Concepts and Models

In order to choose the appropriate computational approaches and experimental methods for this study, it is important to understand the general principles of drug repositioning including the relationships between drugs, targets and diseases, their interactions and associations.

2.4.1 On-Target Drug Repositioning

It is known as “New target for known compound” paradigm (Figure 2.2). It is defined as investigating new biochemical pathways for possible targets to a known molecule (drug) [31].

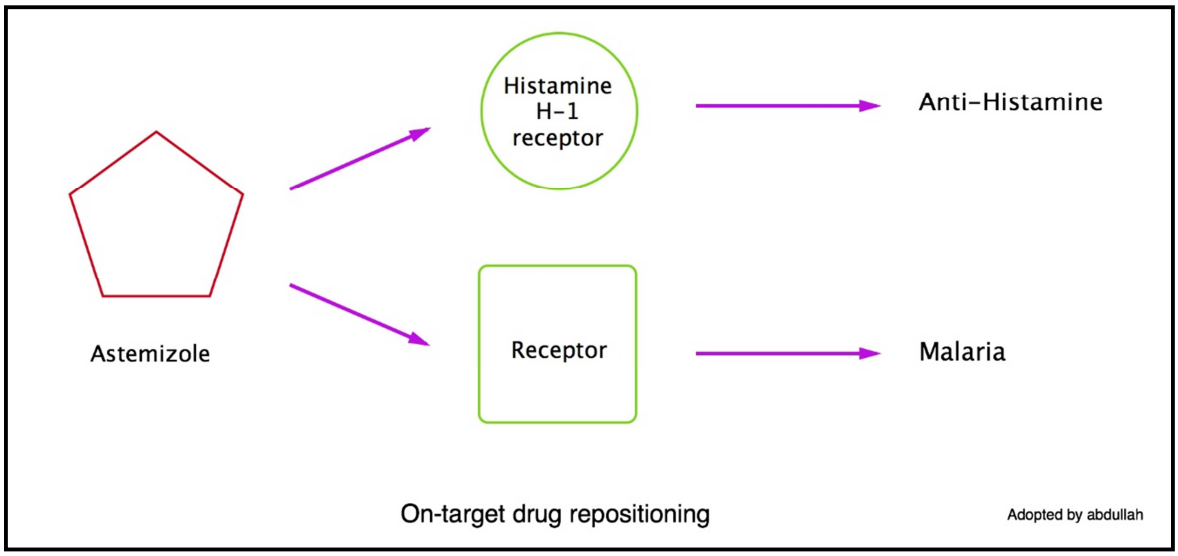


Figure 2.2 Schematic drawing explaining the on-target drug repositioning model. The drug Astemizole, as an example drug, interacts with two different targets (receptors) affecting two different pathways [31]

2.4.2 Off-Target Drug Repositioning

It is known as “new indication for known target” paradigm [31]. It can also be explained as using the same drug for two different biochemical pathways (diseases) which share the same drug target (Figure 2.3).

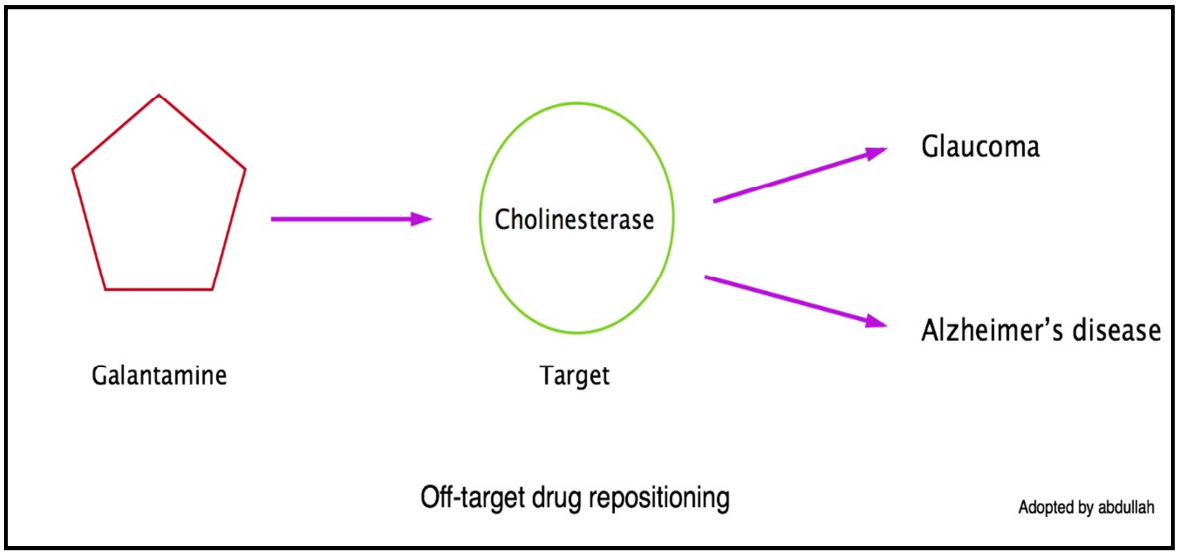


Figure 2.3 Schematic drawing explaining the off-target drug repositioning model. Two different diseases share a single target (cholinesterase) in their pathways. Therefore, Galantamine which is originally designed to treat Glaucoma by targeting the cholinesterase can also be used against Alzheimer [31]

2.4.3 Models of Drug-Target Interaction

These models represent the different strategies in which drugs and targets could possibly interact with each other. Targets might consist of the different forms of biomolecules, such as protein, DNA, enzyme and etc. The interaction can be drug-drug similarity–based interaction, target-target similarity-based interaction, or a combination of both [32].

2.4.3.1 **Triad-Based Model**

This model relies on two inverse strategies to predict the link between drugs and targets:

Target-target similarity triad model: As its name implies, this model involves three elements. If two targets are similar to each other, then they are highly likely to interact with the same drug (Figure 2.4a).

Drug-drug similarity triad model: If two drugs are similar to each other, they are expected to interact with the same target (Figure 2.4b)

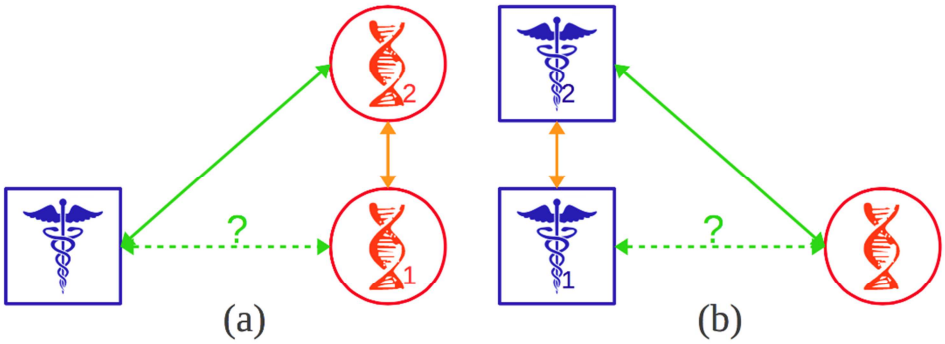


Figure 2.4 The Triad-Based model. (a) Target 1 resembles Target 2. Accordingly, if a known drug interacts with target 2 then the same drug will probably interact with target 1. (b) Drugs 1 and 2 are similar to each other. Accordingly, if drug 2 interacts with a target then drug 1 will possibly interact with same target [32]

2.4.3.2 **Tetrad-Based Model**

If two drugs are similar to each other and two targets are similar to each other, then knowing that the first drug is interacting with the first target suggests that the second drug would probably interact with the second target [32] (Figure 2.5).

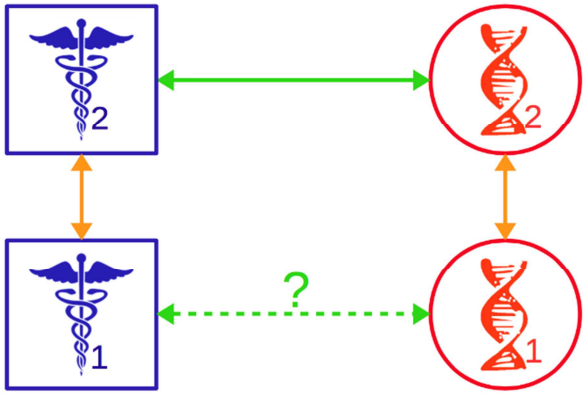


Figure 2.5 The Tetrad-Based model. If drug 2 interacts with target 2, drug 1 resembles drug 2 and target 1 resembles target 2, then drug 1 would presumably interact with target 1 [32]

2.5 **Methods**

The increasing complexity and size of data about diseases and drugs mechanism of action are reflected in the increasing number of drug repositioning methods and techniques. Estimating the amount of available biological and pharmaceutical data and information is an essential step to decide the best methodology to be used for drug repositioning [33].

Drug repositioning researchers have developed numerous methods to study the available data about drugs and diseases. These methods can be broadly divided into two categories: experimental and computational. I will explain briefly these methods with examples.

2.5.1 Screening Methods

Screening methods are either phenotypic screening (*in vivo* and *in vitro* HTS/HCS screening) or by using of FDA off label method (clinical decisions). Both of these methods do not include or depend on biological or pharmaceutical information. Therefore, these methods can not help to determine the mechanism of action for any drug.

These methods are generally considered as of low complexity and simplicity. However, some of the most famous examples of repositioning such as rituximab (for breast cancer), sildenafil (for erectile dysfunction), and HDECC inhibitors (for lung cancer cells) have been discovered and tested serendipitously using these screening methods [34, 35].

The screening method is very flexible giving the ability to test any drug without the need for any prior knowledge about it. This ease of use was the reason behind the discovery of 28 molecules out of 75 approved for clinical usage between 1999 and 2008 [33].

2.5.2 Target-Based Methods

Two examples of target-based methods are *screening* and *cheminformatics*. The screening method consists of *in vitro* and *in vivo* investigations of drugs for correlation to specific biomarker or protein. The cheminformatics method consists of docking, *in*

silico screening, and ligands-based analysis of different compounds from drug libraries [36].

The more information is available about the targets which are involved or directly linked to a disease mechanism, the higher the probability of finding useful drugs is. Without the need for any extra information researchers can screen many compounds just by using known chemical structure information of drugs, ligands, and target 3D structure [36, 37].

Many pharmaceutical companies use these methods for either finding new drugs or repositioning existing drugs. For example, Melior Discovery, an *in vivo* pharmacology company has recently found a new indication of MLR-1023 for diabetes [33].

2.5.3 Information-Based Methods

Information-based methods refer to the use of bioinformatics tools in order to systematically predict a drug-target interaction, predict unknown mechanism of action of a drug, or discover unknown drug-drug similarities for drug repositioning.

Bioinformatics tools and system biology techniques utilizes the available information about drugs, targets, disease profiles, drug-target networks, disease networks (diseasome), chemical structure similarities between drugs, similarity of side effects, and signaling pathways for the purpose of repositioning drugs [33].

In contrast with previous methods, information-based methods use prior knowledge to start from known associations to identify unknown ones, whilst other methods start without any information or with little information about the targets. As an example; with chemical structure information, Simvastatin and Ketoconazole were repositioned for breast cancer. Moreover, with available pathway information a new indication for skin cancer was found by repositioning of Vismodegib [33].

Blatt J and Corey SJ have used these methods of drug repositioning and succeeded in identifying additional drugs for pediatric hematology oncology and for pediatrics generally [38]. In 2015, McCabe *et al* published a research paper on the importance of knowledge-based methods for repositioning compounds toward blood cancer treatment with special concern to dosage and toxicology [39].

2.5.4 Genomic-Based Methods

A large number of diseases are of genetic origin (e.g. autosomal recessive disorders) or are linked to genetic factors that play an important role in the disease etiology. Identifying the genetic causes of a disease on the molecular level is a key factor in the process of finding a treatment for the disease. Genomic techniques have reached an advanced level nowadays. For instance, microarray and next generation sequencing techniques generate a huge amount of disease related data. Genomic-based drug repositioning methods depend on the outcome of these techniques to identify new disease networks and to explore new relations between targets and drugs.

Additionally, the rise of open source genomics databases speeds up the studies of drug repositioning. Some of the examples of these databases are: CMAP Connectivity Map [40, 41], NCBI-GEO [100], SRA Sequence Read Archive [101], and CCLE Cancer Cell Line Encyclopedia [42].

Gene expression data analysis [43], genome-wide association studies [44,45], gene set enrichment analysis (GSEA), microRNAs signatures analysis, and comparative genomic hybridization data measurement, all of these approaches can be modeled to build a conceptual framework for general genome-based drug repositioning [46].

In the case of complex multifactorial disorders with a genetic component such as cancer, the genomic approaches will help in linking these disorders to other diseases through their genetic profiles allowing them to benefit from the same drug. For example, *Imatinib*, an inhibitor originally developed for BCR-ABL fusion protein in chronic myelogenous leukemia (CML), was repositioned afterwards for gastrointestinal stromal tumor (GIST) [46].

2.5.5 Biochemical Pathway-Based Methods

In addition to genomic data, protein-protein interaction networks, and the known metabolic and signaling pathways, can be utilized to build another mode of drug repositioning which depends mainly on signal transduction data.

This method recreates novel disease-specific pathways which provide key targets for drug repositioning. It can help in mining the general signaling networks to scale them down from enormous number of proteins to a specific network with few proteins [33].

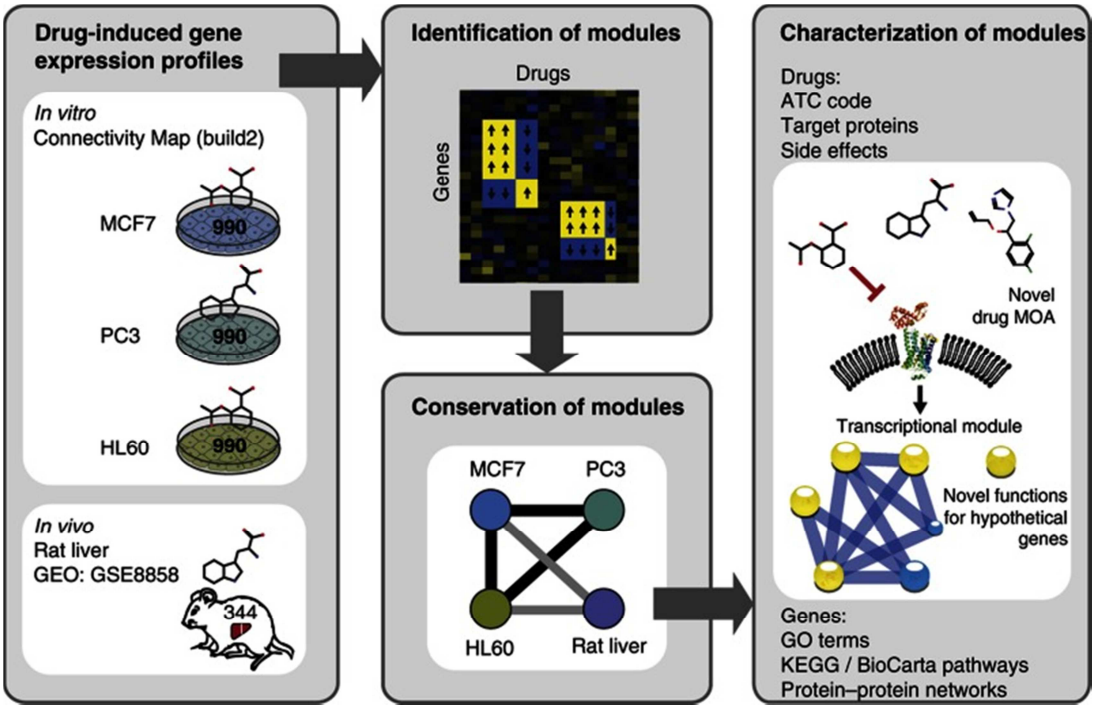
In 2013, Zhao *et al*, [47] developed a computational model to explore specific signaling pathways allowing for the discovery of unknown connections between targets and diseases, and novel mechanisms for specific cancer subtypes. They identified a new type of signaling pathways, called cancer signaling bridges (CSB), which holds great promises for sourcing and facilitating systematic drug repositioning. In addition to that, they established what they called “individualized signaling network”, which showed a new perspective to deal with the complexity of three metastases of breast cancer [47].

2.5.6 Targeted Mechanism-Based Methods

This is a very sophisticated methodology, but it is not well established yet. Several research groups are working to develop the ultimate configuration for this method.

The biggest challenge that is confronted by this method is to identify, in addition to the general mechanisms related to drugs or diseases, the mechanism of drug action on the treatment of specific diseases. To address these challenges, systems biology and network biology approaches are applied to define unknown drug action mechanisms depending on the integration of protein functional networks, signaling cascades, and all omics data [34].

Iskar *et al* [48], in their approaches found 10 novel regulators of cellular cholesterol homeostasis which provided a starting point for drug repositioning. Figure 2.6 explains the workflow of the approach they developed, starting with:(1) Identification of drug-induced modules in human cell lines and rat liver, the data received from two resources: (i) the CMap, and (ii) DrugMatrix, (2) conservation of drug-induced modules across cell types and organisms, (3) characterization of gene and drug members of drug-induced modules, (4) functional discovery within drug-induced modules, (5) rich source provided by drug-induced transcriptional will lead to drug repositioning.



2.6 Drug-induced transcriptional modules method fro drug repositioning and functional understanding [48]

Over all, choosing between the methods mentioned above to start a drug repositioning study depends on the available data. Accordingly, selection of the research method can be prioritized as in Table 2.4 [34].

Table 2.4 Prioritizing drug repositioning methods according to the available data

Available data	**Options**
Little information available for the disease	Phenotypic screening FDA off-label
One protein biomarker for the disease	Target-based methods Knowledge-based methods
More disease information available: disease pathways data, disease omics data, etc.	Knowledge-based methods Genomic-based methods
Treatment omics data (omics data generated from drug treatment)	Genomic-based methods Targeted mechanism-based methods

2.6 **Computational Approaches**

I will concentrate more on this section and explain it in more detail. This section, especially, the approach of data mining using machine learning algorithms is the most related section to my thesis research.

The main aim of drug repositioning is to establish a link between a disease and a drug. Every computational approach tries to build a bridge between different biomedical concepts or any other concepts to prove the link.

2.6.1 Chemical Similarity Approaches

Molecular similarity or chemical structure similarity concepts are very applicable in drug repositioning. The logical principle for this concept is called: *similar property principle* [49]. This principle can be summarized as structural similarity yields to functional similarity.

Similar property principle is derived from a known classification model used in biological science and chemical engineering, **Quantitative Structure-Activity Relationship model** (QSAR model), which suppose a relationship between chemical structures and biological activity [50].

To compute the structural similarity of two chemicals there is a collection of methodologies that can be used such as clustering algorithms and fingerprints [50]. One of the most common measures of structural similarity depending on the fingerprints approach is the Tanimoto (or jaccard) coefficient T, where two structures are usually considered similar if T> 0.85 [51].

In 2006, Noeske *et al*. published one of the most interesting clustering approaches. They used an unsupervised machine learning algorithm (self-organizing map) with topological pharmacophore descriptor (CATS) to predict the interaction mechanism of mGluR antagonists with several receptors [52].

Keiser *et al.* (2009), developed another method called similarity ensemble approach [53], grouping the ligands according to their targets binding partners and predicting thousands of unanticipated associations based on the chemical similarities between drugs and ligands. They used a statistical model to calculate the possibility of a molecule to bind to a target depending on the shared chemical features between the molecule and known target ligands, see figure 2.7.

In summary, similarity principle is defined as drugs with similar structures have similar biological activity. For example, Drug A (which binds to H1 histamine receptor) in Figure 2.7 share some similarity with drug B (which binds to serotonin receptor 5A). This similarity suggests that Drug A could bind to serotonin receptor 5A and generate same activity.

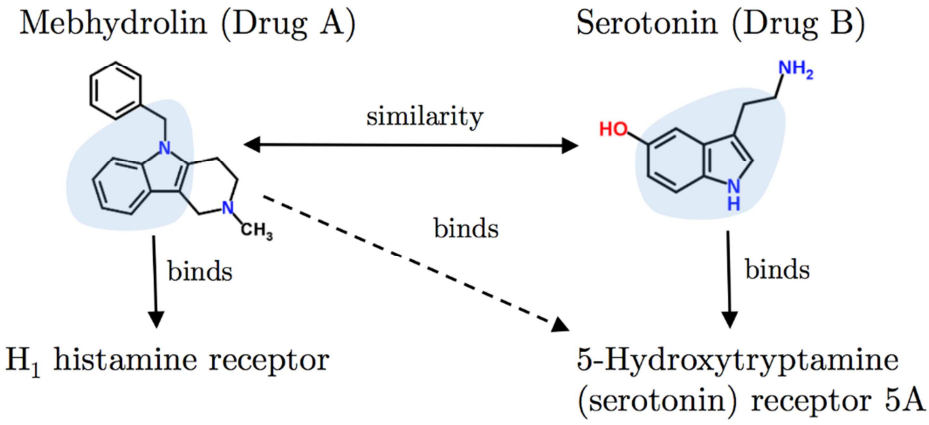


Figure 2.7 Similarity principle [54]

2.6.2 Gene Expression Approaches

The Connectivity Map (CMap) project is a powerful source of information for a variety of research studies. It depends mainly on gene-expression profiles to find new connections between diseases and drugs that share a mechanism of action or chemicals and physiological processes [40].

The differential gene expression is a characteristic of a molecular phenomenon known as the gene expression signature. Genes go through states of over and under expression in response to the different conditions. These states of expression are estimated by the relative numbers of the transcribed messenger RNA (mRNA) molecules for each gene.

The idea behind the CMap states that the action of a drug is measured and then linked to the gene expression signature it creates when administered into a biological system. The data is freely available and can help to perform various types of analyses, such as the identification of the molecular mechanism of a drug. There are some cases in which known drugs that are used for different clinical indications showed similar gene expression signatures. Therefore, it is important to keep in mind when doing similar studies that further validation experiments and tests are required.

Lorio *et al,* (2010) developed an automated approach to exploit similarities in gene expression profiles by using network theory concepts [54]. They built a web-based tool called Mantra 2.0 [102] to analyze the Mode of Action (MoA) of new drugs based on network theory statistics of gene expression data.

Figure 2.8 shows the principle of gene expression-based similarity for drug repositioning and drug-disease association. The gene expression data from the CMap are compared, genes in green color are up-regulated, and genes in red are down-regulated. The data provides a signature which can relate drugs based on their functional aspect. For instance, drug X and Y are considered similar because they affected a significant number of genes in the same manner.

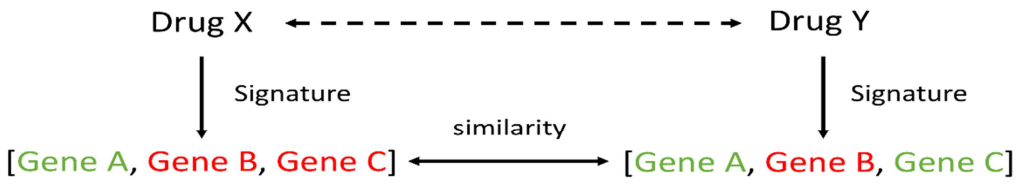


Figure 2.8 Gene expression signature-based drug repositioning [54]

2.6.3 Molecular Docking Approaches

Molecular docking approaches which utilize similar binding sites have become one of the most important tools of drug discovery [55]. Knowing that different proteins might possess similar binding sites, it is reasonable to conclude that similar binding sites most likely bind to the same ligands.

These kinds of approaches shed the light on the protein-drug interaction space, which helps to better understand drug modes of action and can also help in reducing drug doses. The growing amount of data in this field is an advantage that will support the optimization of drugs to gain higher selectivity and thus reduce side effects.

The promiscuity of drugs empowers the ability of one drug to bind to multiple distinct targets. Figure 2.9 shows the process of drug repositioning using binding site similarity. A and B are proteins which were aligned (as we see on C) because of their binding site similarity. Due to the known binding between A and D, it is suggested that D might as well bind with B [56].

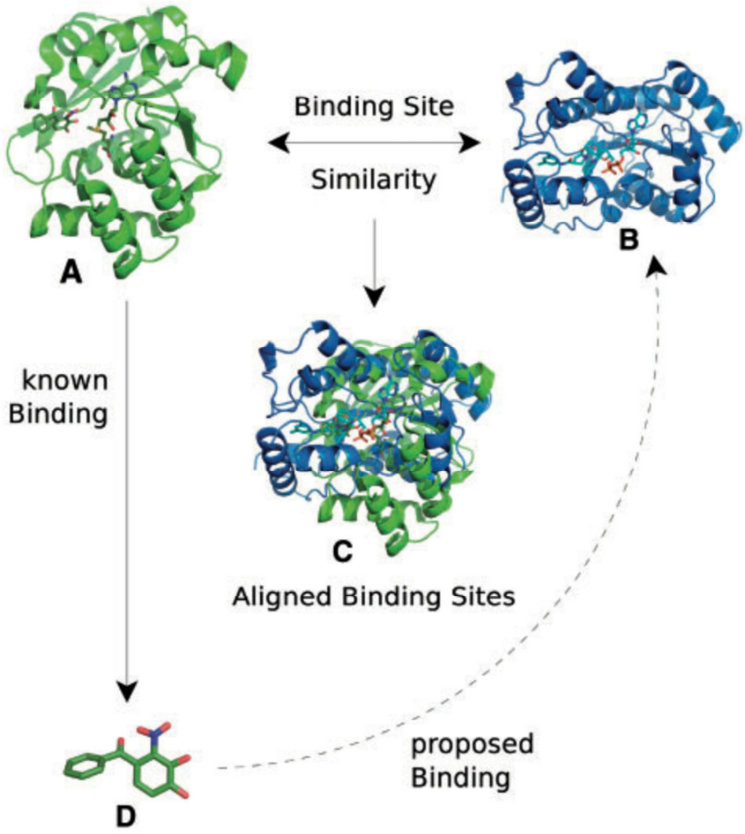


Figure 2.9 Molecular docking concept [57]

In a computational approach to perform molecular docking studies several algorithms were developed to increase the precision and accuracy of the technique [56]. A list of these algorithms with their functions and characteristics is provided in table 2.5 below.

Table 2.5 Molecular docking algorithms

Algorithm	**Function**	**Characteristic**
Matching algorithm (MA)	Based on molecular shape map a ligand into active site of a protein in terms of shape features and chemical information	Geometry-based, suitable to VS and database enrichment for its high speed
Incremental construction (IC)	Put the ligand into an active site in a fragmental and incremental fashion	Fragment-based and docking incrementally
Multiple copy simultaneous search (MCSS)	Randomly placed 1000-5000 copies of a functional group, in the binding site of interest and subjected to simultaneous energy minimization and/or quenched molecular dynamics	Fragment-based methods for *de novo* design

	in the forcefield of the protein	
Monte carlo (MC)	Generate poses of the ligand through bond rotation, rigid-body translation or rotation	Stochastic search
Genetic algorithm (GA)	The mutation and crossover are genetic operators affect the genes (which is binary string encoded to represent a degrees of freedom of the ligand), the result is a new ligand structure.	Stochastic search
Molecular dynamics	Powerful simulation method, in the context of docking, by moving each atom separately in the field of the rest atoms, and represent the flexibility of both the ligand and protein	For further refinement after docking

Molecular docking is considered as an effective method to represent the physical interaction between a drug and a protein. Despite being far from covering the whole proteome, but scientists from all over the world are working on exploring the molecular docking mechanisms for most of the proteins available on the Protein Data Bank (PDB).

2.6.4 Side Effect Similarity Approaches

The simplest way to reposition a drug is to monitor carefully the side effects that might appear during the clinical trail of a drug, which might provide ideas for targets that can be exploited or diseases to be treated with the same drug. As mentioned previously, Sildenafil and Thalimode are examples for the most successful repositioned drugs which were repositioned after the observation of unexpected side effect.

The concept of side effect observations has become very common in biological studies during the last few years. Consequently, a huge increase in the development of computational methods and tools to study diseases and drugs depending on side effect similarities and associations was recognized. These methods involved mainly two big areas: network theory's concepts and property analysis, and machine learning methods and algorithms.

Kuhn *et al*. (2010), developed a database called SIDER [94], as a source of side effects data, which connects 880 drugs to 1450 side effects [58]. The database serves the purpose of facilitating computational drug discovery and drug target predictions via side

effects data analysis. The drugs were grouped in the database according to their drug class, and shared side effects for each class of drugs are also provided.

The basic principle of this drug repositioning approach, as it is explained in figure 2.10, is that the more side-effects shared by two drugs the stronger the similarity between them [54]. The similarity can be used to reposition either off-target or on-target drugs.

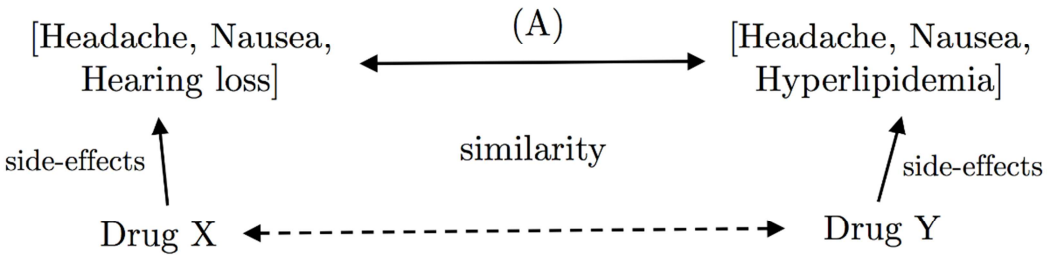


Figure 2.10 Drug repositioning using side-effects similarity concept [54]

2.6.5 Text Mining Approaches

The advances in natural language processing (NLP) technology, have improved the text mining approaches resulting in more precise biomedical data mining studies. One more reason for the high demand and the improvements of text mining approaches is the vast increase in the number of published articles. PubMed, for example, contains more than 20 million articles covering many scientific and medical disciplines [3].

By definition, text mining is a method of textual analysis that transforms the text into significant indexes intended for data extraction and information identification [59]. Text mining in the context of drug repositioning is the attempt to link existing drugs to new diseases by thorough textual mining into the published biomedical abstracts. It is intended to identify alternative drug indications by overlapping or leveraging publicly available information resources and mechanism of action representations.

Barcante *et al.* (2015) developed a drug repositioning approach based on text mining which consists of three distinct phases [60]. Phase 1: setting a programming framework which manifest the terms that will be used to search and recover abstracts in articles downloaded from PubMed. Phase 2: extracting protein names from full text articles that are previously selected, and searching for proteins with similar structure or function in the biological databases. Phase 3: suggest inputs for the repositioning of drugs.

Another approach was proposed by Patchala and Jegga [61] in 2015, in which they built a statistical topic model based on the Unified Medical Language System (UMLS) concepts which can be found in the disease and drug related abstracts in MEDLINE. This approach can be divided into four steps. Step 1: collection of drug and disease related abstracts from MEDLINE. Step 2: using MetaMap to map semantically the extracted disease and drug related abstracts. Step 3: building a topic model to determine the number of inherent topics in the dataset and to calculate the highest likelihood value for the trained model. Step 4: computing the differences between the topic distributions in the selected disease and drug profiles using Kullbak-Leibler divergence, see Figure 2.11.

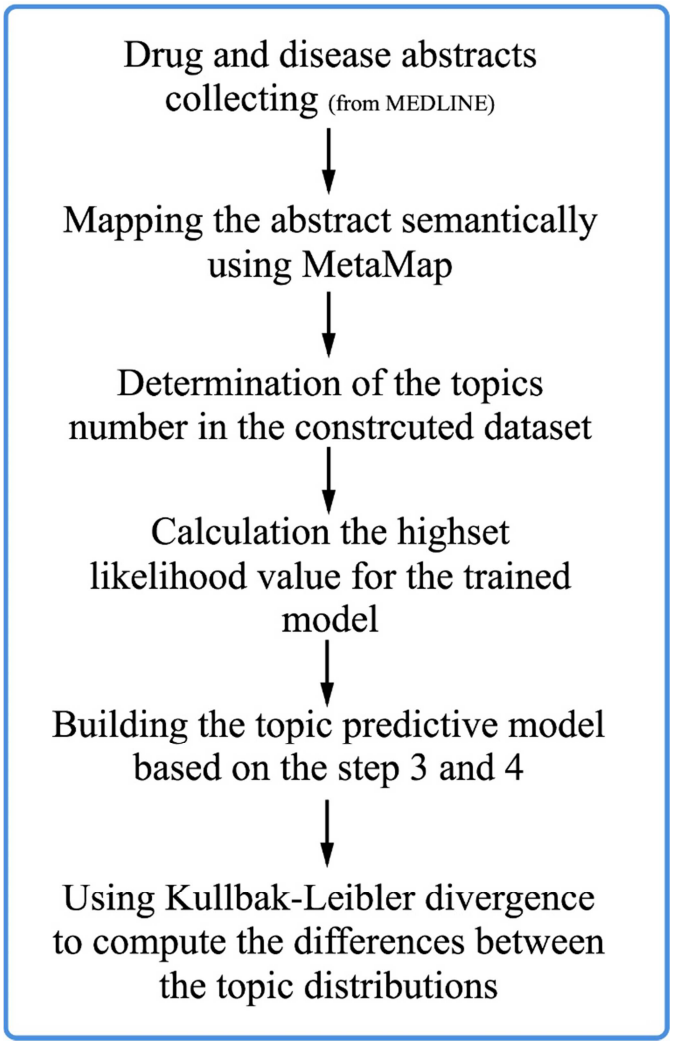


Figure 2.11 Text mining approach based on topic modelling [61]

2.6.6 Machine Learning Approaches

Machine learning is the science of constructing and exploring algorithms that can auto-learn and perform predictions out of a given data [62]. In biomedical studies, it is perfectly feasible to use a combination of the biomedical descriptors such as chemical structure similarity, side effect associations, and protein-protein interactions (PPI), to train a machine learning algorithm and then generate predictions out of the statistical model.

Recently, machine learning has been applied in almost every aspect of biomedical research making great improvements. Figure 2.12, shows a classification of the topics where machine learning methods are applied from system biology to function and structure prediction; it is everywhere [63].

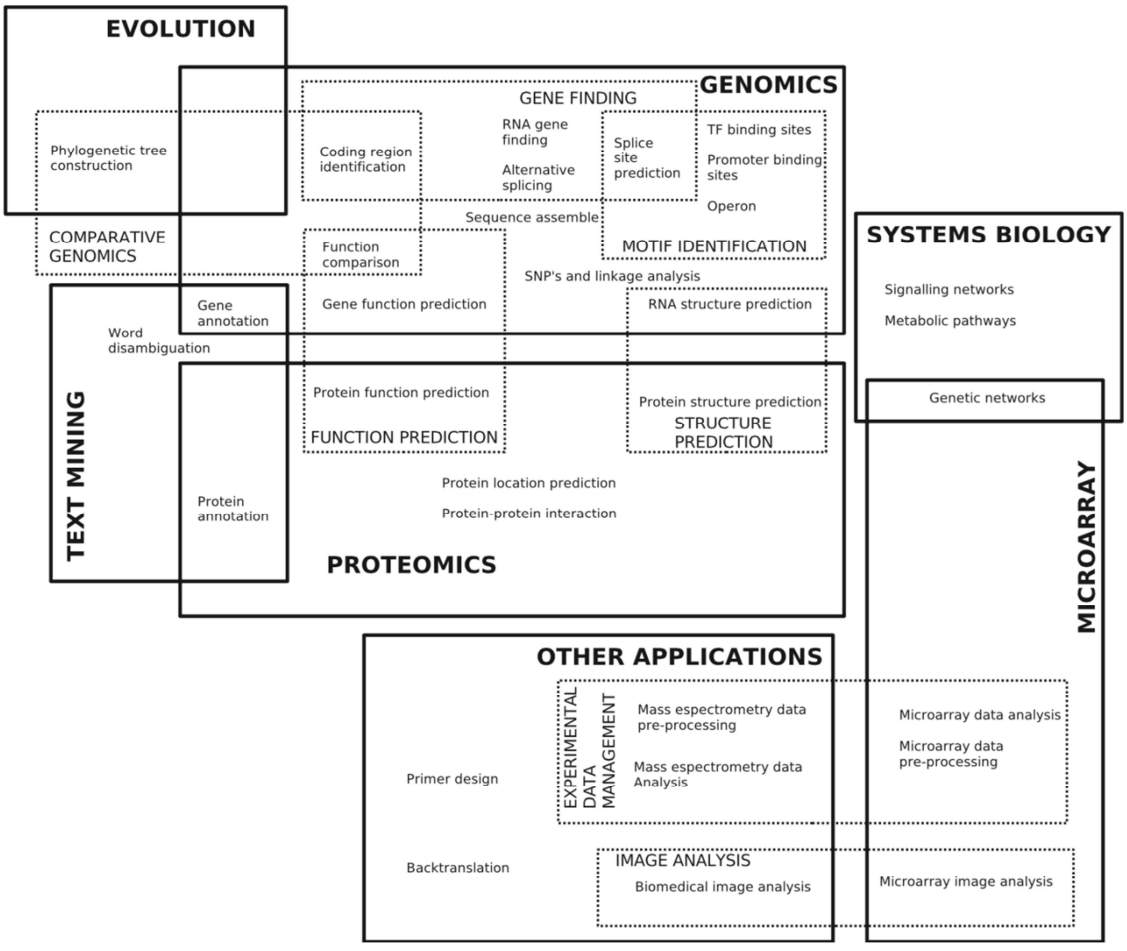


Figure 2.12 Application of machine learning in biomedical topics [63]

Drug repositioning is also an area of research where machine learning approaches yielded several important studies: predicting new associations for unknown drug-disease interactions, using many features or descriptors such as shared targets, side effect similarity, chemical structure similarity, and genetic variations similarity. Figure 2.13 explains the concept of the machine learning approach [54].

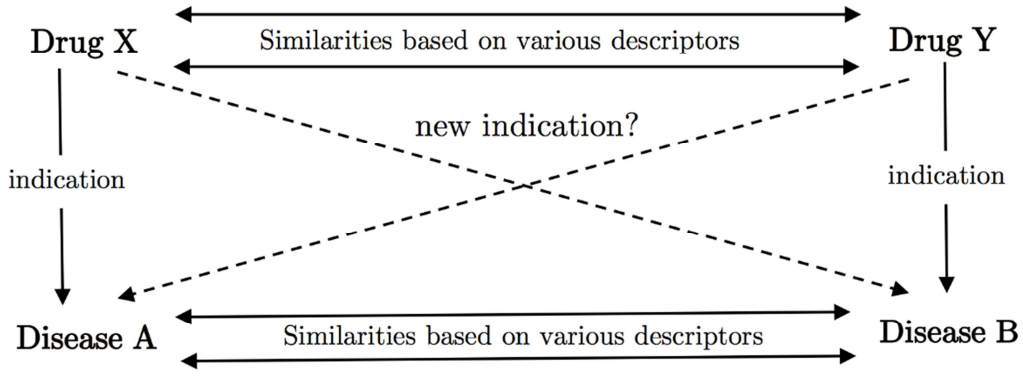


Figure 2.13 The concept of machine learning approach. It uses different combinations of drug-drug similarity and disease-disease similarity to predict new indications [54]

Napolitan *et al* (2013), proposed a machine learning approach by integration and prediction from three types of data: 1) the similarity in chemical structures of drugs, 2) proximity of targets in protein-protein interaction network, 3) correlation of gene expression patterns, based on building a classification algorithm which classifies drugs according to their therapeutic uses [64]. Figure 2.14 demonstrates the workflow of the analysis using this machine learning approach [64].

PREDICT is another example where machine learning algorithms that were designed by Gottlieb *et al* (2011) [65]. It depends on creating a drug-drug similarity and disease-disease similarity matrix to predict a new drug indication according to its similarity to a known drug using a classification scoring system. The method is performed in three steps: 1) construction of drug-drug and disease-disease measures, 2) exploiting these similarity measures to construct classification rule, 3) application of the classifier to predict new associations.

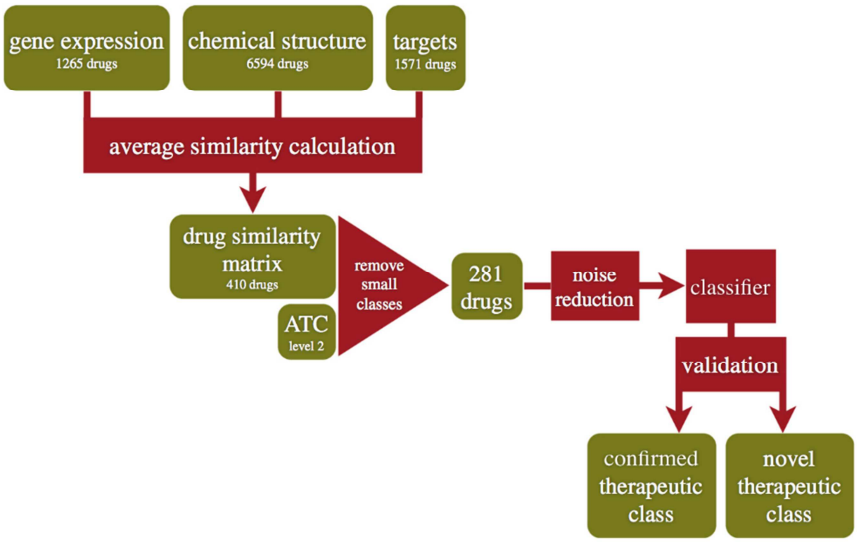


Figure 2.14 Workflow of one machine learning approach example: red boxes are the process; green boxes are the data [64]

2.7 Data sources for drug repositioning

It is highly recommended in any drug repositioning process, particularly computational approaches, to create a list of data sources to extract the required information from it. Computational scientists in biomedical research have access to a wide range of information sources across multidisciplinary fields.

In order to generate a solid model for data analysis, computational drug discovery requires all sorts of clinical, chemical and biological data sources. Data sources that are used for drug repositioning can be divided into three categories: chemical data (cheminformatics) such as chemical structure databases, biological data (Bioinformatics) such as pathways information databases, and literature or textual data like PubMed and MEDLINE, see table 2.6 [33].

Table 2.6 Databases for drug repositioning studies

Field	**Databases**	**Website**
Chemical structure	PubChem	http://pubchem.ncbi.nlm.nih.gov/
	Drugbank	http://www.drugbank.ca/
	Therapeutic Target Database	http://bidd.nus.edu.sg/group/TTD/ttd.asp
	Collaborative Drug	https://www.collaborativedrug.com/
	PharmGKB	http://www.pharmgkb.org/
	ChemSpider	http://www.chemspider.com/

	ChemFrog	http://www.chemfrog.com/
	Chemicalize	http://www.chemicalize.org/
Drug-target information	Drugbank	http://www.drugbank.ca/
	SuperTarget	http://bioinf-apache.charite.de/supertarget_v2/
	BindingDB	http://www.bindingdb.org/bind/index.jsp
	Chemical-Protein Interactions	http://stitch.embl.de/
Literature houses and research tools	PubMed	http://www.ncbi.nlm.nih.gov/pubmed
	MEDLINE	http://www.nlm.nih.gov/bsd/pmresources.html
	Google Scholar	https://scholar.google.com.tr/
Target 3D structure	Protein Data Bank	http://www.rcsb.org/pdb/home/home.do
	OCA	http://oca.weizmann.ac.il/oca-bin/ocamain
	OPM	http://opm.phar.umich.edu/
	Proteopedia	http://proteopedia.org/wiki/index.php/Main_Page
	TOPSAN	http://www.topsan.org/
Side effects	SIDER	http://sideeffects.embl.de/
	FAERS	http://www.fda.gov/Drugs/
	Clinicalrial.gov	http://clinicaltrials.gov/
Molecular omics data	NCBI-GEO	http://www.ncbi.nlm.nih.gov/geo/
	Sequence Read Archive	http://www.ncbi.nlm.nih.gov/Traces/sra/
	Stanford Microarray Database	http://smd.princeton.edu/
	ArrayExpress	http://www.ebi.ac.uk/arrayexpress/
	Princeton University MicroArray database	https://puma.princeton.edu/
	CellMiner	http://discover.nci.nih.gov/cellminer/
	Oncomine	https://www.oncomine.org/resource/login.html
	Cancer Cell Line Encyclopedia	http://www.broadinstitute.org/ccle/home
Genetic data	dbSNP	http://www.ncbi.nlm.nih.gov/projects/SNP/
	OMIM	http://www.omim.org/
Pathway information	NCI-PID	http://pid.nci.nih.gov/
	KEGG	http://www.genome.jp/kegg/

Pathway information	BioCarta	http://www.biocarta.com/
	Reactome	http://www.reactome.org/
	PathwayCommons	http://www.pathwaycommons.org/about/
Drug omics data	Connectivity Map	http://www.broadinstitute.org/cmap/
	CCLE	http://www.broadinstitute.org/ccle/home
	NCBI-GEO	http://www.ncbi.nlm.nih.gov/geo/
	SRA	http://www.ncbi.nlm.nih.gov/Traces/sra/
Protein interaction information	HPRD	http://www.hprd.org/
	BioGRID	http://thebiogrid.org/
	STRING	http://string-db.org/
	MIPS	http://mips.helmholtz-muenchen.de/proj/ppi/
	IntAct	http://www.ebi.ac.uk/intact/
	DIP	http://dip.doe-mbi.ucla.edu/dip/Main.cgi

2.8 Examples of Notable Databases

I will briefly describe four databases that represent four concepts of drug repositioning. These databases were developed to be a source of information and a computational tool which provide the capacity of analyzing its own data content. The concepts represented here are: analyzing complex networks based on probability theory, analyzing networks based on similarities, molecular docking, and disease-disease similarities.

2.8.1 Pharm DB

Pharm DB [103] is a Korean platform based on network structural and topological properties analysis [9]. This platform aims to establish a linkage between drugs, proteins, diseases and side effects, and to discover unknown associations in various complex networks. It targets four kinds of nodes: drugs or chemical molecules, drug targets which can be protein or DNA, diseases, and side effects.

To do so, they constructed an algorithm called Shared Neighborhood Scoring (SNS), to predict any kind of relationships between the four kinds of nodes. The algorithm depends on probability theory to evaluate and calculate the connections between two nodes (which can be drug and protein for example). It sums up what they called “Shared Nodes Count”, the number of shared nodes, and “Shared Nodes Weight”, the product of each weight of direct or indirect links bridging the two end nodes. The represented

nodes can be one of these combinations: drug-disease, drug-drug, drug-protein, drug-side effect, disease-protein, and protein-protein see Figure 2.15.

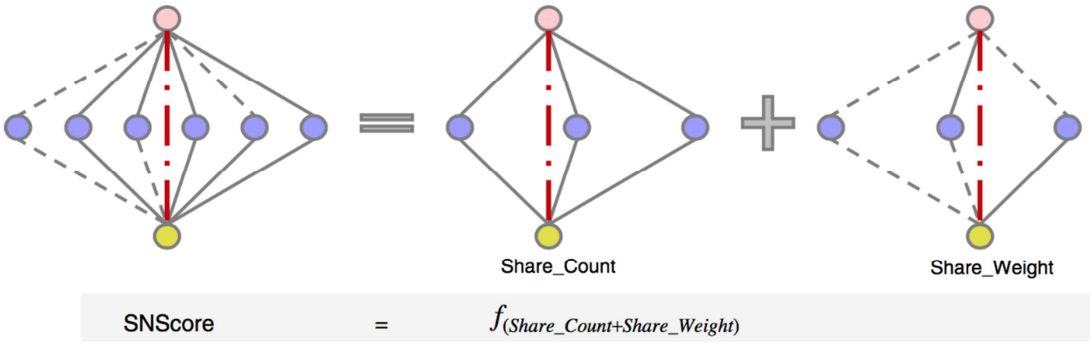


Figure 2.15 Shared Neighborhood Scoring (SNS) Algorithm [9]

The pharm DB platform has three interfaces: web browser for data collection, phExplorer for data visualization, and BioMart for predicting the shortest path between two nodes. Figure 2.16 shows an overview of the pharm DB platform [9].

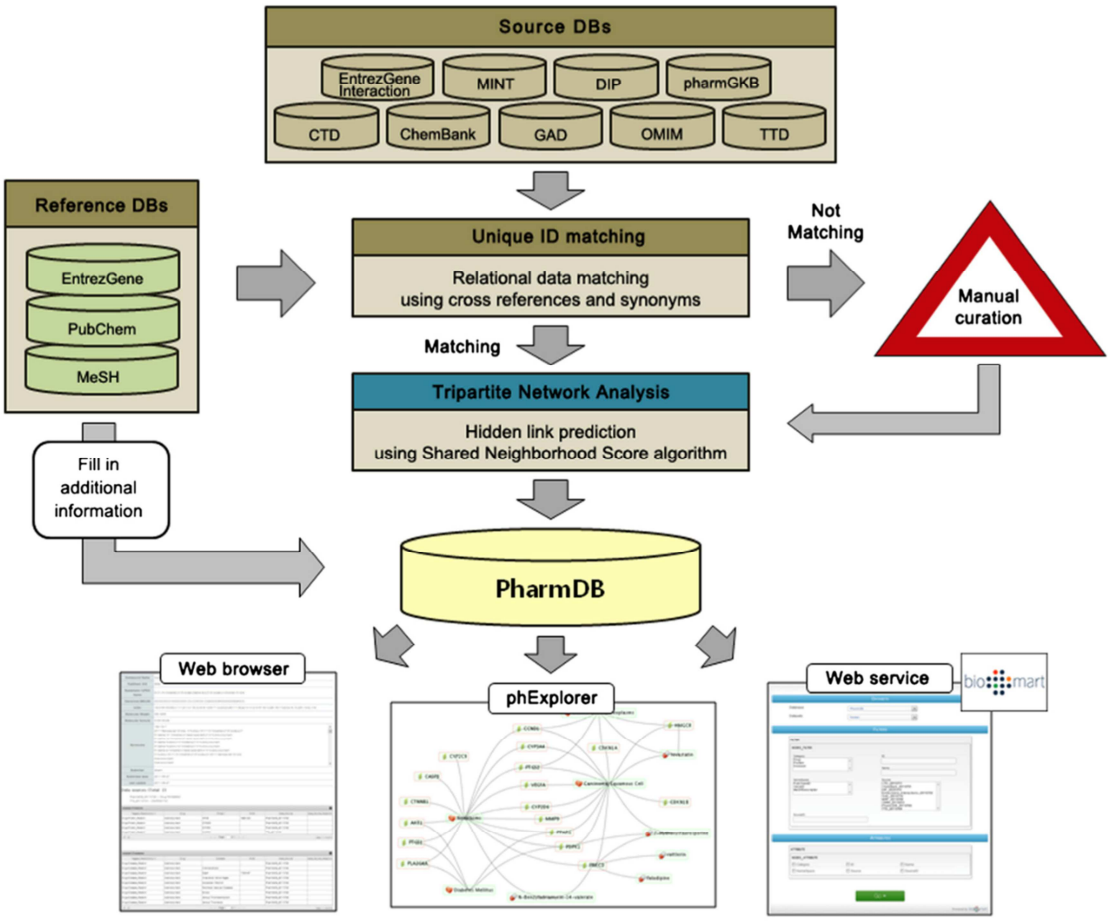


Figure 2.16 Overview of pharm DB which showing the nine data sources in top, with terms references to overlap the synonyms, the shared neighborhood scoring algorithm, and the three component of the interface of the platform [9]

2.8.2 Promiscuous

Promiscuous, the term is derived from the word promiscuity which means the ability to have several partners or connections in the same time. It is a web-based tool [104] and network-focused database which contain three types of entities: proteins, drugs and side-effect data. The analysis is performed on combinations of the different entities interaction possibilities: drug-drug, drug-protein, protein-protein, and drug-side effect. The interaction combinations are subjected to a rule-based classification algorithm which classifies the drug-drug structural similarities, the networks of protein-protein interaction and its distances, and so on for drugs side effects.

As shown on Figure 2.17, promiscuous has five interfaces: 1) **data visualizer** which can represent the entities as nodes and can analyze the nodes and their links, 2) **KEGG pathway mapping** which allows users to explore the drugs or proteins involved in some signaling and metabolic pathways, 3) **full text search** that can view any ID information of a drug or a disease in any context in detail, 4) **relation viewer** which provides the information of the relationships between entities in detail, 5) **pin board interface** which facilitates searching and enables the saving of searched information to make it available anytime.

Performing a search in promiscuous, based on drug similarity features, the database developers found an important connection between two known drugs. Memantine, a drug prescribed for dementia in Alzheimer's disease patients, and Amantadine, an anti-Parkinson drug, were found to share the same target (NMDA glutamate receptor) in their mechanism of action [66].

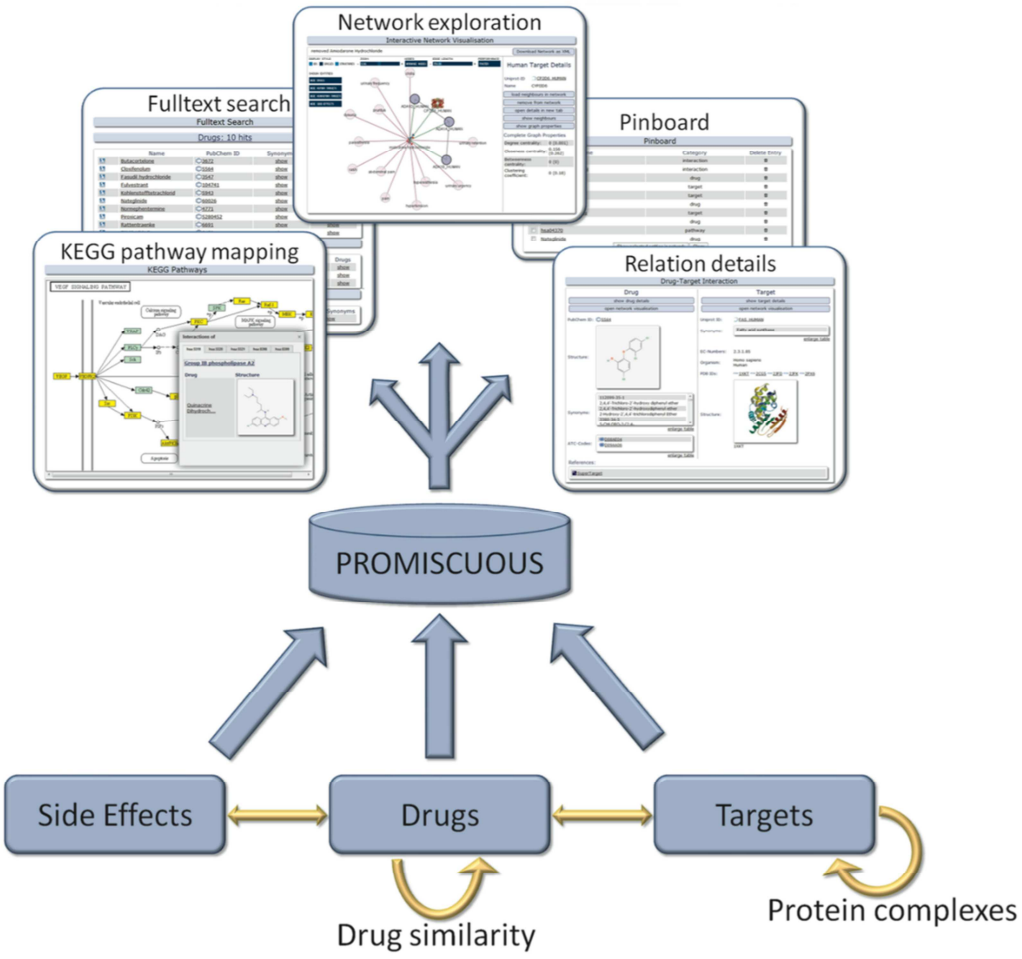


Figure 2.17 Promiscuous schematic illustration [66]

2.8.3 DRAR-CPI

DRAR-CPI [93] is a server that contains a library designed to find candidates for drug repositioning based on two elements: Chemical-protein interactome (CPI) and adverse drug reaction (ADR). The server works in parallel with DOCK, a software engineered in the 1980s to predict the binding modes of small molecules. DOCK uses geometric algorithms to calculate the probability of docking score [67].

The DRAR-CPI server has created an in silico association data using the DOCK software, and established a library of 385 target proteins and 254 small molecules. The drugs included in the library have an extensive description of their indications and ADRs which allows for the prediction of repositioning candidates based on the association scores between library molecules.

Figure 2.18 bellow, illustrates the main stages in the DRAR-CPI server workflow. The workflow starts by uploading a drug file in the format of mol or smiles. Then, the DOCK will hybridize the drug with all the targets in the library and will calculate the binding affinity scores. The binding scores of the uploaded drug will then be compared to the drugs existing on the server library resulting in negative and positive association scores representing a weak or strong correlation respectively [68].

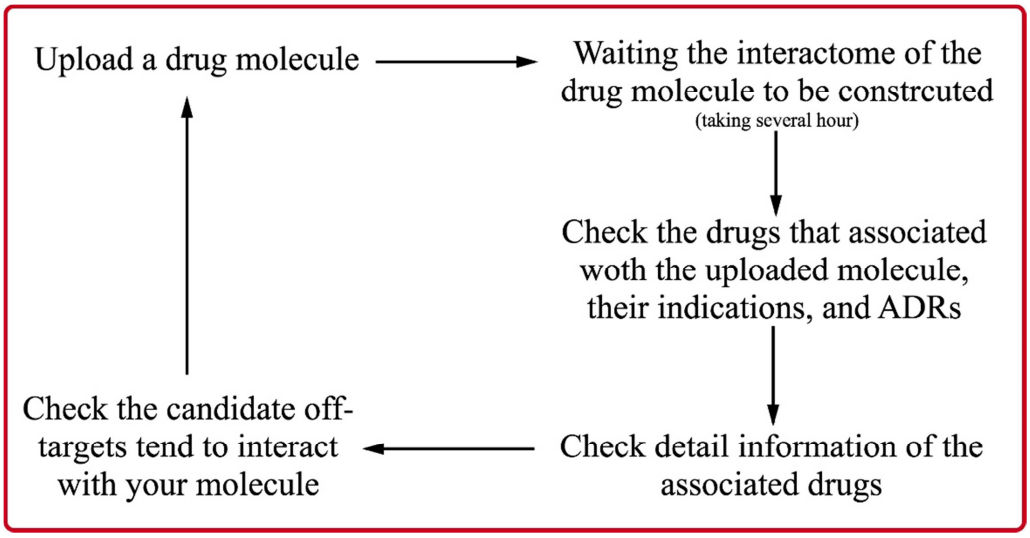


Figure 2.18 procedure of DRAR-CPI work [93]

2.8.4 Disease-Connect

Disease-connect [105] is a web-hosted server aims to classify diseases using a completely new strategy, trying to solve the problem of traditional classification method

which groups diseases according to their clinical symptoms and phenotypic traits. Diseases with similar clinical presentations and completely different etiologies will be grouped together which will mislead the drug designing process. Disease-Connect overcome this problem by classifying diseases according to their molecular etiology not their phenotypic presentations. This classification method will provide a substantial data source for the repositioning of existing drugs toward the treatment of many diseases [69].

The user interface on the Disease-Connect platform provides the possibility to search for a gene, a disease or a comparison between two diseases. The data outcome of diseases correlation search includes: disease-related gene expression, disease-related microRNA expression, disease-related SNP, disease-drug, disease-comorbidity, disease-gene relationships from Gene RIF database, disease-gene relationships from literature corpus mining, and disease-gene relationships from OMIM.

Figure 2.19 represents the Disease-Connect operational network for the analysis of the correlation between two or more diseases. It combines the disease-gene associations from three different databases. Then, it enriches this combination by appending the disease-drug relations to it. In addition to that, it calculates the *p*-value for the number of genes shared between two diseases to assess the significance of the relation between these diseases.

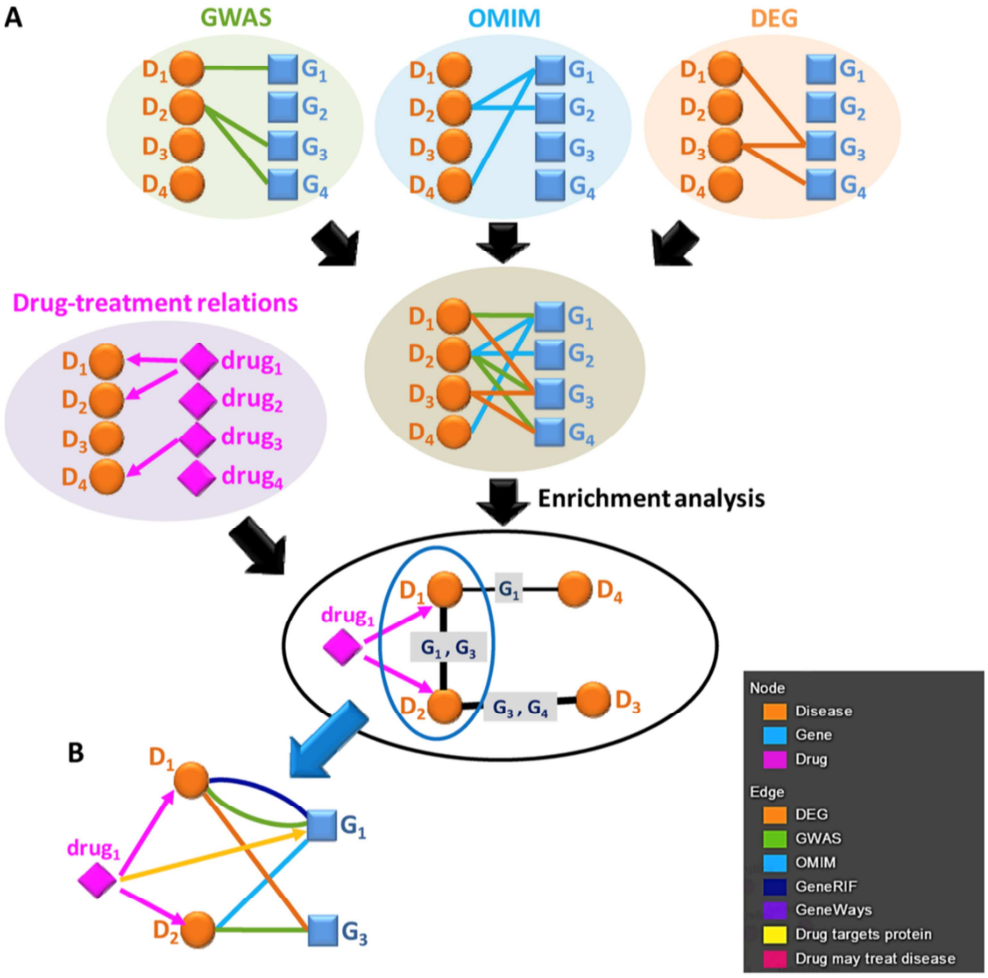


Figure 2.19 Overview of Disease-Connect server [69]

CHAPTER 3

METHODS AND MATERIALS

3.1 Why Machine Learning?

Machine learning (ML) is a subfield of artificial intelligence, which depends on intelligent algorithms that enables the machine to learn and predict from a provided data. In the data perspective, machine learning is a type of data mining that improve the program's own ability to extract and recognize the patterns inside the data.

I decided, in my thesis, to apply machine learning techniques to predict unknown drug-disease associations (patterns) in order to reposition existing drugs to multiple diseases. The process is initiated by feeding the programming language (Python in my case) with data about drugs, diseases, and proteins resulting in calculated associations between the provided data inputs. The reasons behind the use of machine learning can be summarized in the following points:

1. Big biomedical data available online
2. ML is fast and time saving
3. ML can lead to highly accurate results
4. Can incorporate many descriptors, compared to one descriptor in other repositioning methods
5. The existence of many drug interaction models to work with

Generally, the process of machine learning can be divided into several main steps [63]:

1. Collecting the data (in our case, collecting structure, activity, phenotype data for drugs, and the data that will characterize the diseases).
2. Integration and merging the different data sources into only one format, in purpose of resolute and detect of outliers and inconsistencies, and solve it.

3. 90% of the time in machine learning projects is spent on selection, cleaning, transforming, and correcting the uncorrected data. (step 2 and 3 known as data preparation).

4. Taking the objectives of the study into account in order to choose the most suitable analysis method for the data (data mining step), and selecting the model of learning according to the encountered problem. In machine learning there are two main types of learning:

A. Supervised learning: (I will explain it in more detail, because most of my work is based on the supervised learning), but in general, supervised learning is an algorithm that uses a known dataset (called the training data) to make predictions, the input and desired output value are determined.

B. Unsupervised learning: is an exploratory data analysis to discover hidden patterns or grouping in data, without labeled responses, leaving the algorithm to find structures on the data on its own.

5. The last step after obtaining the predicted model from the data is the testing, evaluation, and interpretation of the model in the biological and statistical perspectives.

Yaser S. Abu-mostafa, professor of machine learning at California institute of technology, and author of classic book in the field titled “Learning from data”, said: in any learning problem and machine learning task you need three components to apply machine learning techniques to the problem you are facing [70]:

1. A pattern exists, and we have to find it (unknown drug-disease associations in our case)

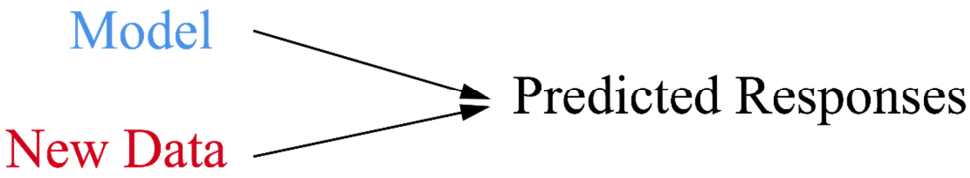
2. We cannot pin it down mathematically

3. We have data on it

3.1.1 Supervised Learning

As I have mentioned, the thesis is based on supervised learning framework, which can be achieved by two main steps: 1. Taking a known data set as an input, and known responses (desired labels) to the data as an output to train a model (Figure 3.1), 2. Generating predicted responses using the trained model on a new dataset (Figure 3.2).

Figure 3.1: First step of supervised learning [71]



3.2 Second step of supervised learning [71]

There are two main categories of supervised learning: regression analysis and statistical classification. Here is the explanation of both categories:

1. Regression is the prediction of a changeable and continuous measurement for an observation. Response variables are real numbers.
2. Classification is the process of taking a data set as an input, and assigning a class (labels can be: true or false) from a finite set of classes to the input. Classes are categorical variables [71].

3.2 **Materials**

The essential data involved in my thesis include drug chemical substructures, drug targets information [72], drug side effects [73], disease-related genes and disease-miRNA expressions. To stand on the shoulders of previous research work, I used already constructed data sets from one of the most successful works on systematic drug repositioning. Then, I extracted determined amount of data and integrate it into a proper structure to fit the thesis analysis format, which I will explain in the next section of this chapter.

3.2.1 Sources

The datasets that were used in this thesis have been acquired from multiple databases because individually each database was not comprehensive enough.

1. **DrugBank**: DrugBank is the most powerful tool and comprehensive database for any drug-related data including drug target information, drug chemical structures, drugs side effects data, pharmacogenomics data, drug action pathways information [106]. It is also useful for any *in silico* application such as drug target discovery, rational drug design, drug metabolism prediction, molecular docking [74]. DrugBank contain until now 7,795 (small molecule, experimental, biotech) drugs, 237 unique enzymes (with 3483 drug-enzymes associations), 4,140 unique targets (with 15,376 drug-target associations), 117 unique transporters (with 1769 drug-transporter associations), and 24 unique carriers (with 321 drug-carrier associations) [75].

2. **PubChem**: PubChem [95] is part of the National Center for Biotechnology Information (NCBI) databases. It is an open source of data about biological activities of small molecules, molecular structure and bioassay data. It enables biomedical researchers and medicinal chemists to search, retrieve and analyze the available data. PubChem is divided into three main parts [76]: i) PubChem Substance which contains descriptions about the chemical structures, substructures, and their biological activities, ii) PubChem Compound for searching into the chemical properties of unique compounds or similar action group searching, iii) PubChem BioAssay is a database that provide the biological activity experimental data from different sources.

3. **KEGG**: an abbreviation of Kyoto Encyclopedia of Genes and Genomes [107]. The original purpose of this database was to become a main source of information for the genes and their functions [77]. Nowadays; KEGG is collection of many databases with tools to explore and search in these databases, including: KEGG DRUG, KEGG DISEASE, KEGG PATHWAY, KEGG GENES, KEGG GENOME, KEGG BRITE, KEGG COMPOUND (for small molecules) and KEGG REACTION (for biochemical reactions) and others. In perspective of statistics, KEGG contain: 10,305 drugs, 1,870 drug groups, 17,484 compounds (metabolites and other small molecules), 1,432 human diseases, and 17,855,904 genes (not only human genes) [78].

4.**SIDER**: [93]is a specialized database on drug related side effects data, including normal and placebo drugs. SIDER became a primary source for biomedical research that based on side effects data [91]. The database contains 5,880 side effects, 1,430 drugs, and 140,064 drug-side effect pairs.

3.3 **Methods**

The method goes into many steps: i) drug and disease data acquisition, ii) preprocess, clean and prepare the data, iii) drug and disease feature extraction and build the associations profiles, iv) reduce the feature and select the most predictive features v) building the classification model using support vector machine (SVM) algorithms, vi) test the model into new data and predict new drug-disease associations, see Figure 3.3 which for the workflow of the work.

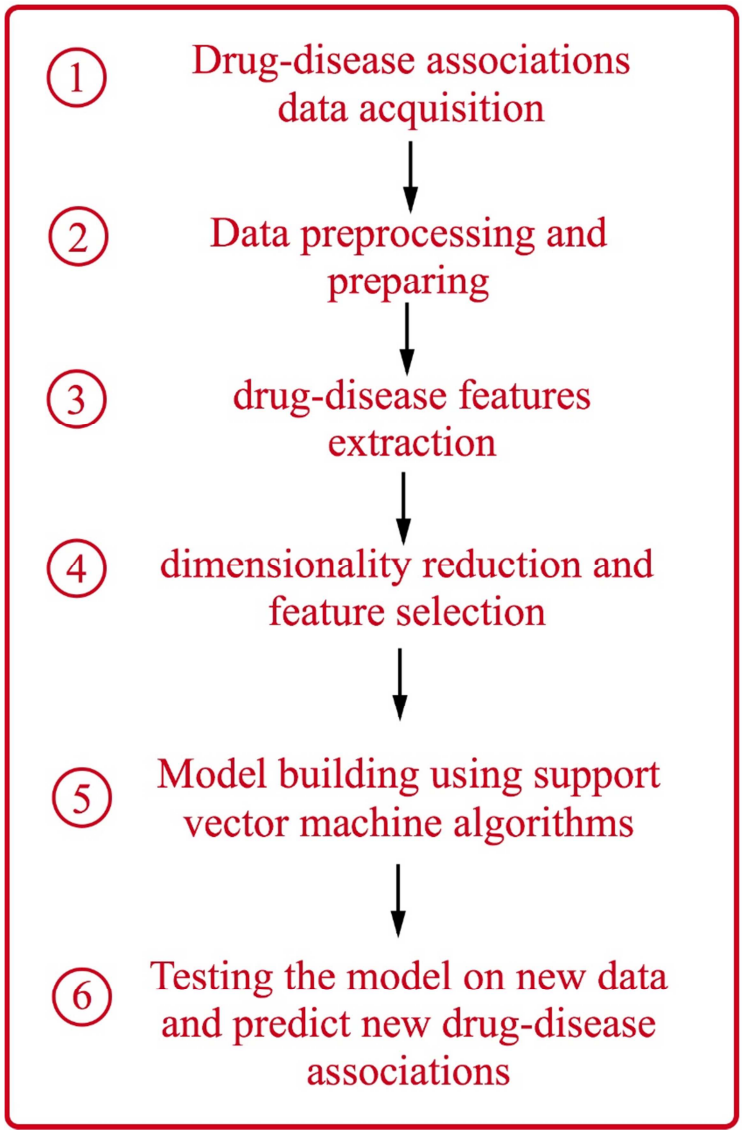


Figure 3.3 Workflow

3.3.1 Feature Types

1. Chemical Structure

The principle of chemical structure property is derived from the medicinal chemistry traditional concept which suggests that similar molecules exert similar biological activities within the biological systems [81]. That similarity can be searched and explored computationally by 3D structure or 2D fingerprint.

To calculate the chemical structure similarity, I used Tanimoto coefficient. In order to understand the Tanimoto coefficient I have first to explain the Jaccard Index or Jaccard similarity coefficient, which was originally coined (1912) by the Swiss scientist Paul Jaccard. Jaccard index is a statistical measurement that compares the similarity and diversity between two data sets, it is a ratio of the intersection between two sets over the union of them [82], which is presented as:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

T.T. Tanimoto developed his similarity and distance index at IBM labs [83], which have the same mathematical model (similarity ratio) of Jaccard Index, but the application proof of concept was different. In the chemical structure similarity context, Tanimoto coefficient calculates the similarity (SIM_{chem}) between two drugs (d_x, d_y) which were presented as a binary fingerprints $f(d_x)$ and $f(d_y)$ to indicate the existence of predefined fragment [84]:

$$SIM_{chem}(d_x, d_y) = \frac{f(d_x) \cdot f(d_y)}{\left|f(d_x)\right| + \left|f(d_y)\right| - f(d_x) \cdot f(d_y)}$$

Where: $f(d_x)$ is the number of fragments of d_x, $f(d_y)$ is the number of fragments of d_y, and the dot is used to produce the number of shared fragments of d_x and d_y.

To compute the calculation of Tanimoto similarity coefficient, PubChem developed a web-based tool known as Score Matrix Service, which enables the user to identify 2D

fingerprint similarity for an uploaded matrix CSV file resulting in scores in the range of [0…100].

2. Side Effect

Side effect means in biological and pharmacological sciences all the effects that appear when trying a molecule or a drug which were unintended originally upon the development of the drug. It is a secondary, unknown, and unexpected effect. These kind of effects which are observed during experimental testing or clinical trials are divided into two types: i) therapeutic effect, which is considered as a positive drug effect which was unknown before, ii) adverse effect which is harmful and might be a lethal effect in some cases.

In both types, the reason behind using side effect data to identify drug-disease associations in purpose of repositioning drugs is that the side effect might highlight and delineate an unobservable mechanism of action for the tested drug [73]. Specific effects are linked to specific pathways, when the tested drug produce a certain side effect which is linked to a known pathway new drug targets can be identified. Finally, drugs that generate similar effects (or side effects) might share some similarity in terms of their targets [73].

3. Drug Target

Identification of new drug targets is an important approach in drug repositioning, and a first step of drug development. If drug repositioning research can help to speed up this step, then the results will speed up all the process of drug production. Drug target also is a core and fundamental property to construct a similarity association between two drugs involving that they share a similar target or a large percent of multi-targets in drug combinations. The targets might be genes, proteins, and enzymes (which have been used) by changing their functions, downregulating their functions, or upregulating their functions. Subsequently, by calculating the shared targets similarity between two drugs (SIM_{target}) D_x and D_y would suggest a functional similarity between them, leading to the assumption that D_x can treat unknown common diseases treated by D_y.

Based on the materials provided earlier, two types of calculations were used to investigate targets similarities, associations and interactions. First: the distance measurement which is used to calculate the closeness between two drug-related gene

pairs [69]. Second: using an already calculated Smith-Waterman sequence alignment scores provided by MimMiner [84], which have been done for all genes and genetic disorders available on OMIM.

4. miRNA

Micro RNA is a short RNA molecule of about 22 nucleotides in length that possess a regulatory function. The main function of miRNA is to downregulate the gene expression at the post-transcriptional level. It is considered as a key element for cellular signaling and cell viability, therefore maintaining the miRNA levels is very critical in most biological systems. The dysregulation of miRNA expression can lead to a wide range of diseases such as cardiovascular diseases and some types of cancer [80].

The many interesting feature of miRNAs, including specific secondary structures and conserved sequences, elected them to be good potential targets for drug design. Therapeutic uses of miRNA are preferred over the use of a mixture of small interfering RNAs (siRNAs) that are specifically designed to bind to specific messenger RNAs (mRNA) leading to their cleavage. Accordingly, specific miRNAs are believed to become the next treatment targets for a majority of diseases [85].

In terms of drug repositioning, most of the data published about miRNAs and their specific features and functions can be viewed in two ways: i) miRNA-drug associations data (available at SM2miR database) which provide experimentally validated effects, and contains (for *homo sapiens*) 161 drugs, 748 miRNAs, and 2307 miRNA-drug associations, ii) miRNA-disease associations data (available at HMDD database) which contains 502 miRNAs, 396 diseases, and 5075 associations.

The potential use of miRNA for drug repositioning rises with the importance of using it as a drug target. Several associations can be built between drugs and diseases based on the affected miRNAs. Knowing the miRNAs involved in the manifestation of certain untreatable diseases and comparing them with miRNAs that are affected by known drugs is a hopeful approach for the repositioning of known drugs toward many diseases.

3.3.2 Machine Learning Format

All the previous works in drug repositioning combines their data using kernel methods to be ready to fuse the data into kernel matrix, which is known as similarity-based

machine learning. I prefer in this thesis to work with traditional machine learning where there are the instances (objects) of interest, features (or attributes) that will be the descriptors of the instances, and the classes (in classification context, where the instances will be: associated 1 class, non-associated 0 class).

Figure 3.4 shows the format of similarity-based machine learning using kernel methods, and Figure 3.5 shows the format of traditional machine learning, where there are the instances (gene as example) and features. In the Table 3.1 I will show the format of this work where it will summarize the instances, features, and classes.

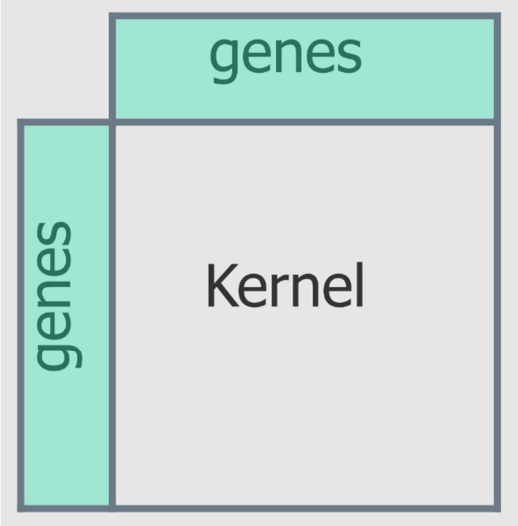


Figure 3.4 Format of the data in the similarity-based machine learning using kernel method [92]

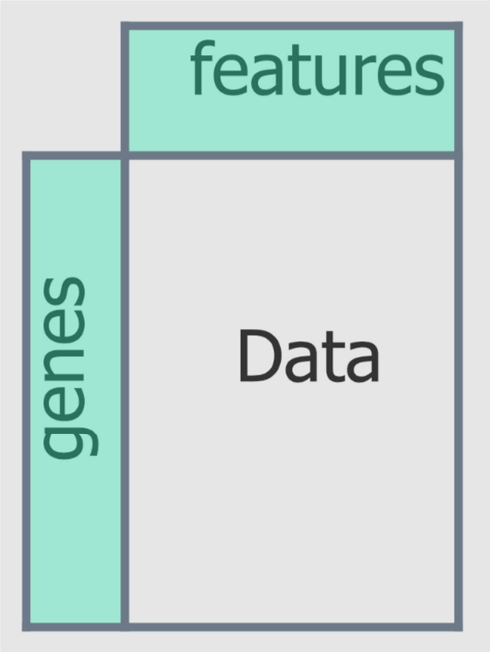


Figure 3.5 Format of the data in the traditional machine learning [92]

Table 3.1 The form of learning: summarizes the instances, features, and classes of this work, where 1 stand for associated drug-disease pairs, and 0 for non-associated drug-disease-pairs

Instance	Feature 1: Chemical structure	Feature 2: Side effect	Feature 3: Drug target	Feature 4: miRNA expression	Feature 5: Disease-related genes	class
Drug1-disease1						1
Drug1-disease2						1
Drug2-disease1						0
Drug2-disease2						0

3.3.3 Dimensionality Reduction

In the preprocessing step of data mining, dimensionality reduction is bringing out the useful part of the data by reducing the number of non-useful or missing features [86]. It is also known that the transformation of data from high-dimensional space to low dimensional space incorporates the most variable and valuable data by removing irrelevant (e.g., near duplicates, poor predictors) and weakly relevant features.

There are two approaches of dimensionality reduction: i) feature selection, and ii) feature extraction. Feature selection techniques used to simplify the predictive model, reduce the data training times, and avoiding the Overfitting while feature extraction, which is very important in image and signal processing is an operation of extraction or identification of meaningful and accurate features inside raw data. See Figure 3.6 to get the idea of feature extraction, and Figure 3.7 which illustrate the whole process and the role of feature extraction and selection (dimensionality reduction) on it.

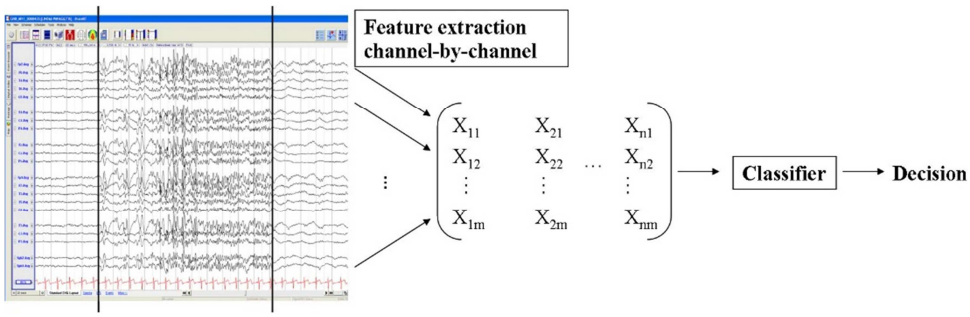


Figure 3.6 Feature extraction [88]

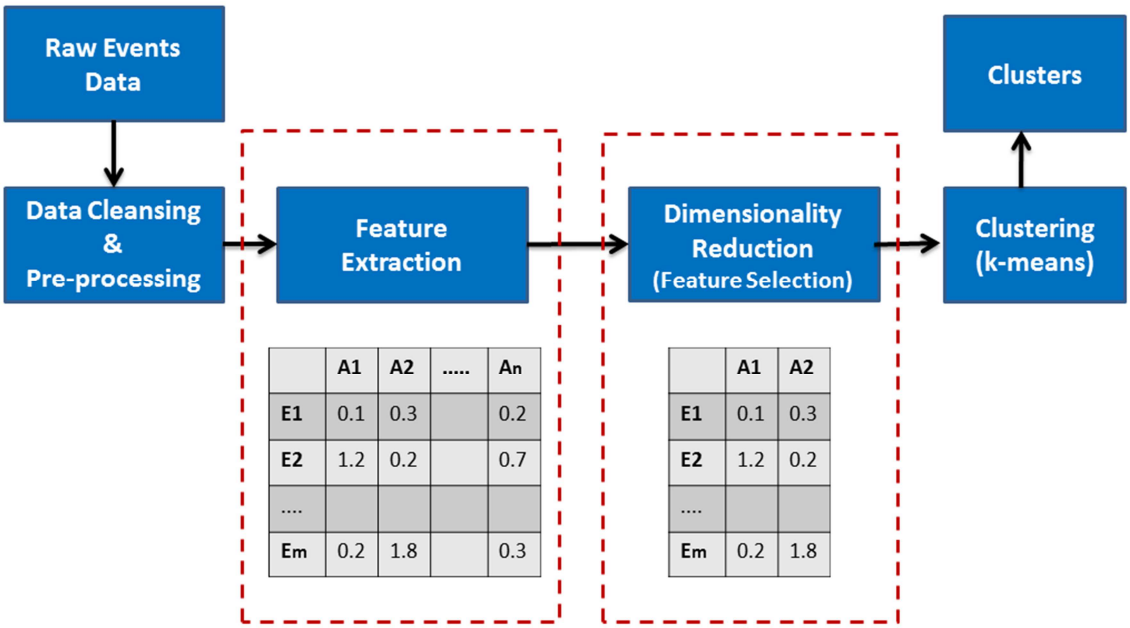


Figure 3.7 Dimensionality reduction in data mining [88]

One of the most important issues in machine learning after overfitting is the curse of dimensionality which dimensionality reduction techniques can help to solve. Curse of dimensionality occurs when adding more features compared to the number of instances leads to poor classifier performance [89].

Figure 3.8 shows the relationship between increasing dimensionality and classifier performance. There is a point called *optimal number of features* where the classifier performs the best until this point with the addition of extra features. However, each additional feature above this point decreases the performance of the classifier.

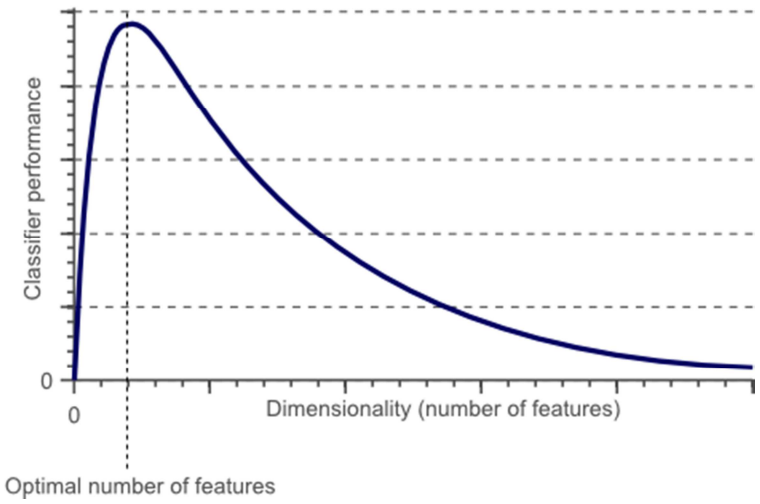


Figure 3.8 Dimensionality and performance [91]

3.3.4 Classification of Dataset

In this thesis, I will use supervised classification framework to solve binary vector-like problem which represent drug-disease pairs association. In addition, I will use support vector machine (SVM) algorithms which belong to the classification category of machine learning, to train a classifier on the data set to build a predictive model that will predict if the drug is associated with the disease or not.

3.3.4.1 **Classification Definition**

In a classification problem, there are a class-like sets of instance, the classes (in our case: associated, and non-associated of drug-disease pairs) are divided based on defined features and classification rules. The role of supervised classification algorithms is to discover the unknown classification rules from the raw data.

In statistical classification, there is a feature binary vector X, which contains two main components: i) predictor variables and ii) class variables C = {0, 1}. The predictive model result from induction of classifiers from the training data which contain a set of *N* observations $D_N = \{(\mathbf{x}^{(1)}, c^{(1)}, \ldots, (\mathbf{x}^{(N)}, c^{(N)})$ [90].

3.3.4.2 **Support Vector Machine (SVM)**

Support vector machine (SVM) is a linear discrimination type of learning, and decision based algorithm for making predictions by classifying the data into classes or several groups. SVMs algorithms generally classify the instances based on a linear function of the features, the following equation describes this linear discrimination (a general linear model) mathematically:

$$F(\mathbf{x}) = w_0 + w_1x_1 + w_2x_2 + \ldots$$

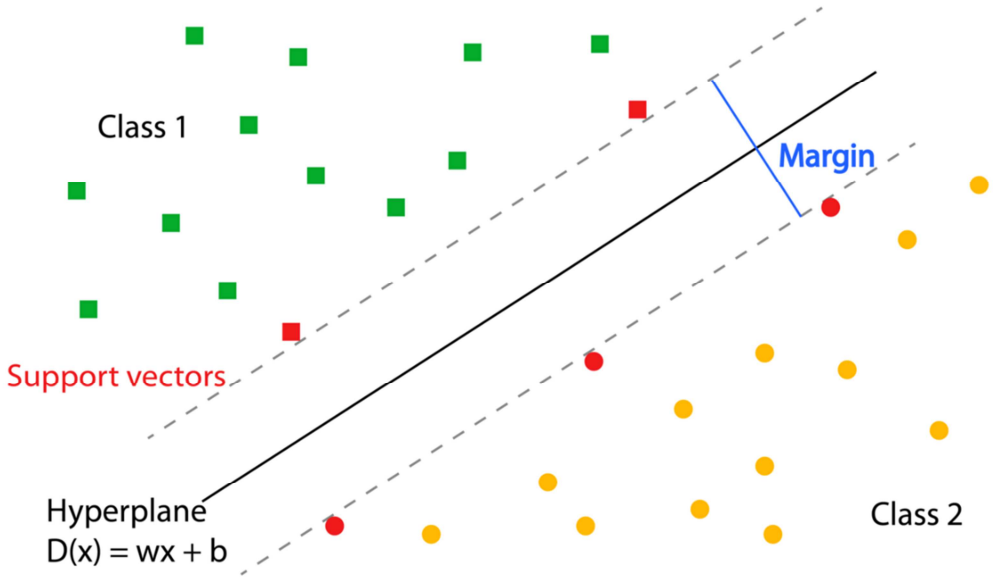
Considering that **x** is feature vector, and x_i is individual features.

Support vector machine algorithms are based on a simple idea: firstly, fitting the widest bar between the classes (which is known as the maximal margin), then separating the classes with a line which will be in the middle of the bar (known as the separating hyperplane) [90].

In other words, support vector machine is based on three basic ideas: i) finding the most proper hyperplane which will separate all points of class one from all points of class two, the most proper hyperplane is the one that has widest margin between the two

classes, ii) the margin is the maximal width that has no inner points between two parallel lines of the hyperplane, iii) support vectors are the closest point to the hyperplane [90].

A simple example is shown in figure 3.9, there are two classes, green squares are data of class one while orange circles are data of class two. The support vector machine algorithm sets up the black line (hyperplane), and margin (blue and dotted gray parallel lines) separating the two classes. In addition to that, support vectors (in red: circles and squares) defines the hyperplane.



3.9 Support vector machine

3.3.4.3 **Two Approach of Classification**

1. Two class classification using (SVM) problem, where I tried to distinguish between two classes of objects. In the thesis context, class 1 (represented as 1) is associated drug-disease pairs, while class 2 (represented as 0) is non-associated drug-disease pairs.

2. One class classification using (SVM) problem, where I tried to describe one class of objects, distinguish it from all other possible objects and find the outliers. I the thesis context class 1 for associated drug-disease pairs, where class -1 is the outliers.

3.3.5 Validating the Classification Model

The main phases in any classification task are: training phase, and testing phase. Training the data set, also known as model building, is using part of the data set to train

the algorithm. The testing step is applying the model to unseen data using known values to compare the values that are predicted with the known values. The reason behind the testing phase is to assess the accuracy of the model predictions.

In this study, as it is usually in model evaluation, I used the accuracy to test the classification performance and examine its ability to construct (predict) the drug-disease associations. Accuracy is defined as the percentage of the closeness of predicted values to the known values. The following equation is used to calculate the accuracy:

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

3.3.6 Scripts and Programming Tools

In this thesis, I used two programming language, R language for data acquisition and preprocessing, and Python for data analysis and machine learning. R language is a powerful language for applied statistics with thousands of packages; see Table 3.2 which contains these packages and their functions. Python is an important language for data science with libraries that makes the data analysis easier; also see the used libraries in Table 3.3.

Table 3.2 R language packages

Package	Function
XLConnect package	Provides comprehensive functionality to read, write and format excel data
Tidyer package	Used for data tidying, reshaping and aggregating
Dplyer package	Used for data manipulation
Openxlsx package	Used for read, write and edit XLSL files

Table 3.3 Python libraries

Library	**Function**
Pandas	Data analysis library
NumPy	Scientific computing library
Scikit-learn	Machine learning library
Matplotlib	2D plotting library
Seaborn	Statistical data visualization library

CHAPTER 4

RESULTS AND DISCUSSION

In this chapter, I will present the hybridized data sets, the dimensionality reduction results, the classification results and the classification accuracy. In addition to that, I will end that with a conclusion based in the results.

4.1 **Hybrid Dataset**

The original size of data I acquired from the databases is shown in Table 4.1. The hybridized data sets made of different combinations of the features that related to the drug-disease associations. Using the original data, I prepared four data sets which contains different types of features, and I chose to work with the fourth data set because it contains all the types of the features and the suitable amount of associations. See Table 4.2 which summarizes these data sets.

Table 4.1 Original size of data

Data	Original size	
Drug-disease known associations	Drug	1066
	Disease	364
	associations	1976
Drug-chemical substructures	Drug	888
	Chemical substructure	881
Drug-side effect	Drug	888
	Side effect	881
Drug-targets	Drug	11771
	Target	11771
Disease-genes	Disease	3306
	Genes	3306
Disease-miRNA expressions	Disease	1207
	miRNA expression	1207

Table 4.2 Hybrid datasets

Data set	Chemical substructure	Side effect	Target	Gene	miRNA	Drug	Disease	Associations
1	✓	✓	✗	✓	✗	109	109	369
2	✓	✓	✗	✓	✓	56	56	926
3	✓	✓	✓	✓	✗	106	106	208
4	✓	✓	✓	✓	✓	55	55	639

Table 4.3 shows information about the number of each type of feature of the selected data sets. In addition, I checked the frequency of the diseases and the drugs, because each drug associated with many diseases, and each disease associated with many drugs. Figure 4.1 contains a plot for disease occurrence, and Figure 4.2 contains a plot for drugs occurrence.

Table 4.3 Information about selected data set

Drug-disease pairs number	Chemical substructure	Side effect	Target	Gene	miRNA	Features total numbers
639	491	817	137	65	137	1647

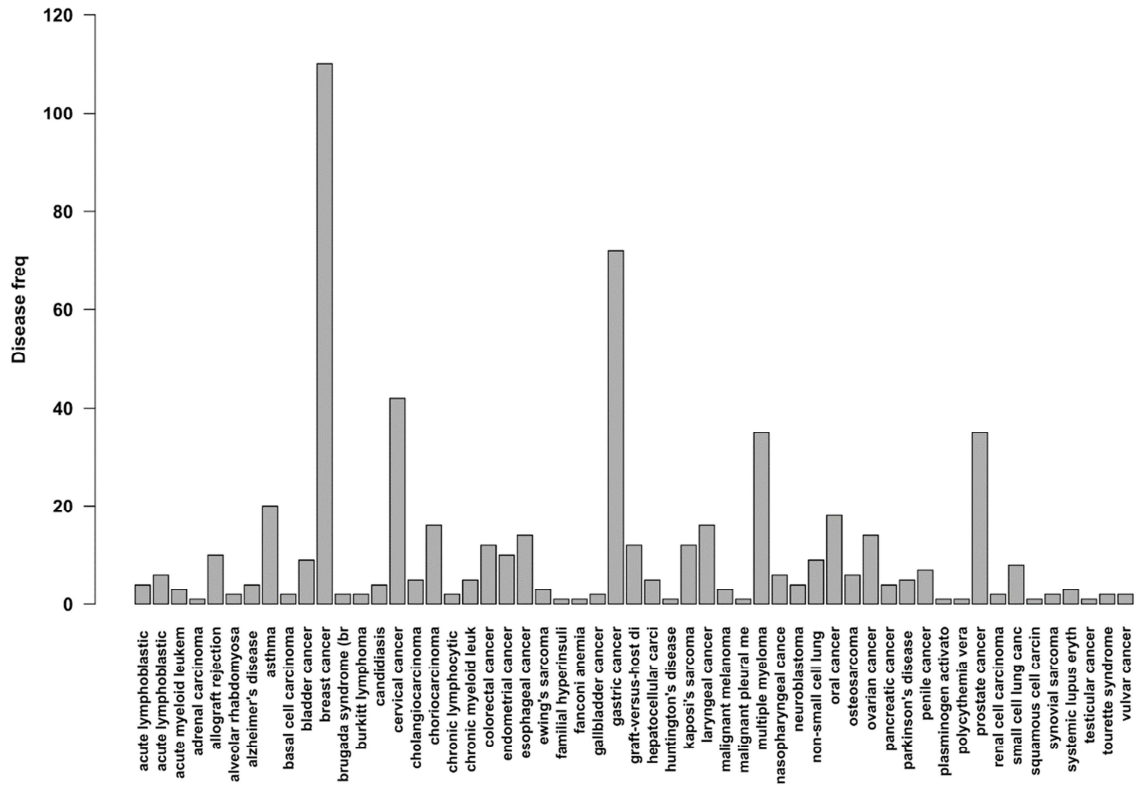


Figure 4.1 Number of the diseases occurrence

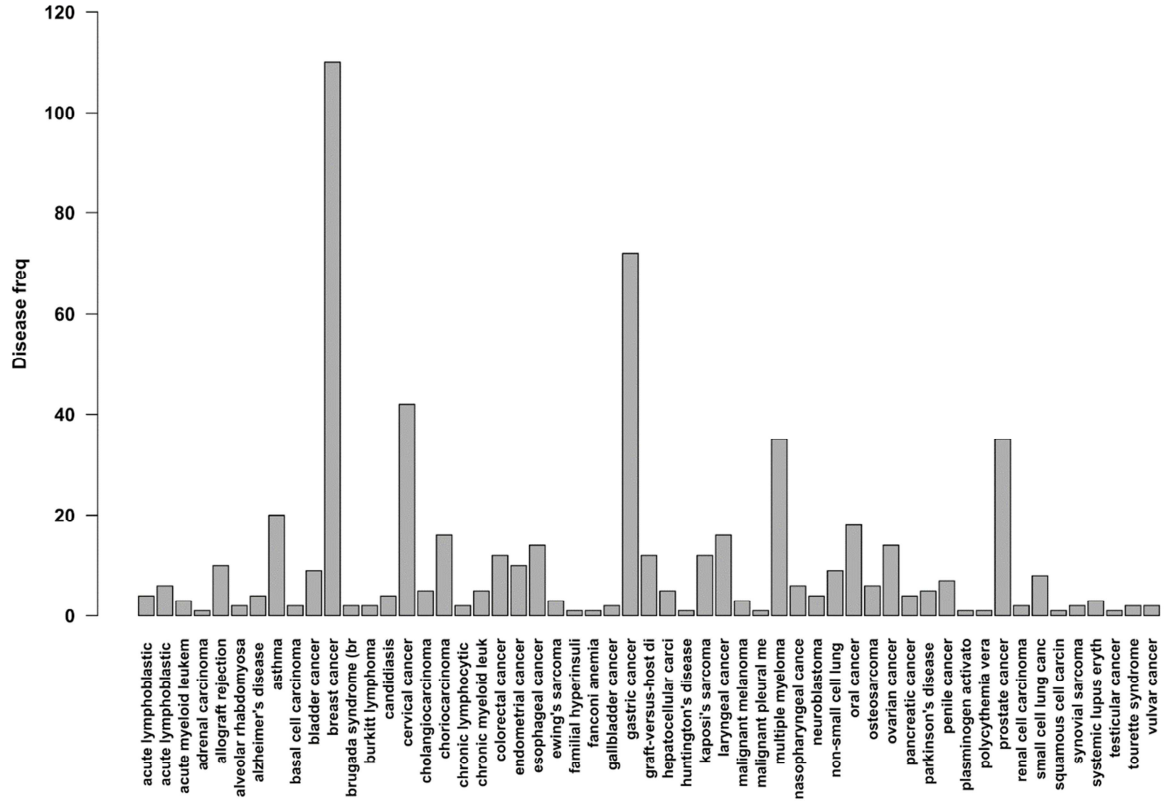


Figure 4.2 Number of the drugs occurrence

4.2 **Basic Statistics**

First of all, I checked the mean and the standard deviation values of the features of the selected data set, to get first hunch for the overall distribution of the entire dataset. Table 4.4 contains the mean value of each feature type and Figure 4.3 shows its chart. Table 4.5 contains the standard deviation value of each feature type and Figure 4.4 shows its chart.

Table 4.4 Mean value of each feature type

Feature	mean
Chemical substructure	0.319308
Side effect	0.139238
Drug target	0.030746
Genes	0.021823
miRNAs	0.039346

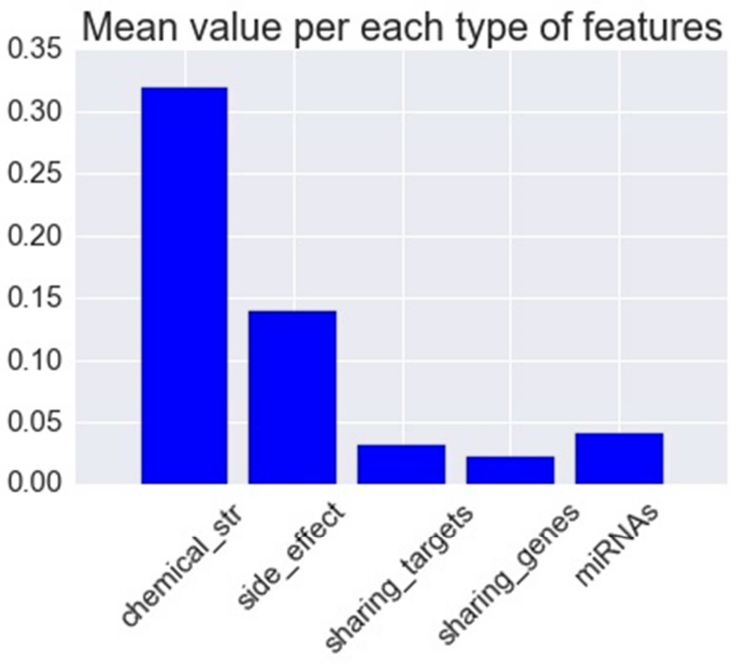


Figure 4.3 Chart for feature types mean

Table 4.5 Standard deviation value of each feature type

Feature	mean
Chemical substructure	0.466209
Side effect	0.346195
Drug target	0.172631
Genes	0.146105
miRNAs	0.194417

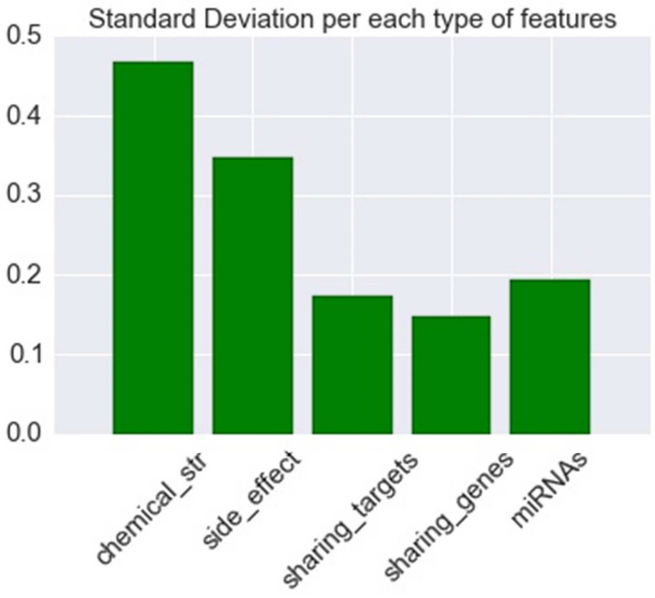


Figure 4.4 Plot Chart for feature types standard deviation

4.3 **Dimensionality Reduction**

1. Dimensionality reduction for two classes classification approach: to prepare the data set for two classes (SVM) and to use dimensionality reduction methods, I needed to split the data set into training and testing set, then to randomize each data set to have positives and negatives. Table 4.6 shows the size of each after splitting, and Table 4.7 the data sets after randomization.

Table 4.6 Training and testing data sets splitting

Data set	Object	Feature
Original data set	639 pairs	1647
Training data set	439 pairs	1647
Testing data set	200 pairs	1647

Table 4.7 data sets before and after randomization

Training data	439 pairs	Before randomization	All positives pairs
	1000 pairs	After randomization	439 positives pairs
			561 negatives pairs
Testing data	200 pairs	Before randomization	All positives pairs
	500 pairs	After randomization	200 positives pairs
			300 negatives pairs

To reduce the features and select the best predictive ones, I used three methods available on Python scikit-learn library for machine learning: i) feature importance using Extra trees classifier, ii) recursive feature elimination (RFE), iii) principal component analysis (PCA). Using each method, I reduced the features into 100 features and 20 features.

i) Feature importance using Extra trees classifier:

A) 100 features approach: Table 4.8 contains the selected 100 features according to this method for each feature type. Figure 4.5 shows the chart of the selected 100 features, where we can see that the biggest number of important features is from miRNA type. Where there is no any chemical substructure or drug target features.

Table 4.8 100 features reduction using Extra trees classifier for two classes SVM approach

Feature type	Amount
Chemical substructure	0
Side effect	1
Drug target	0
Gene	25
miRNA	74

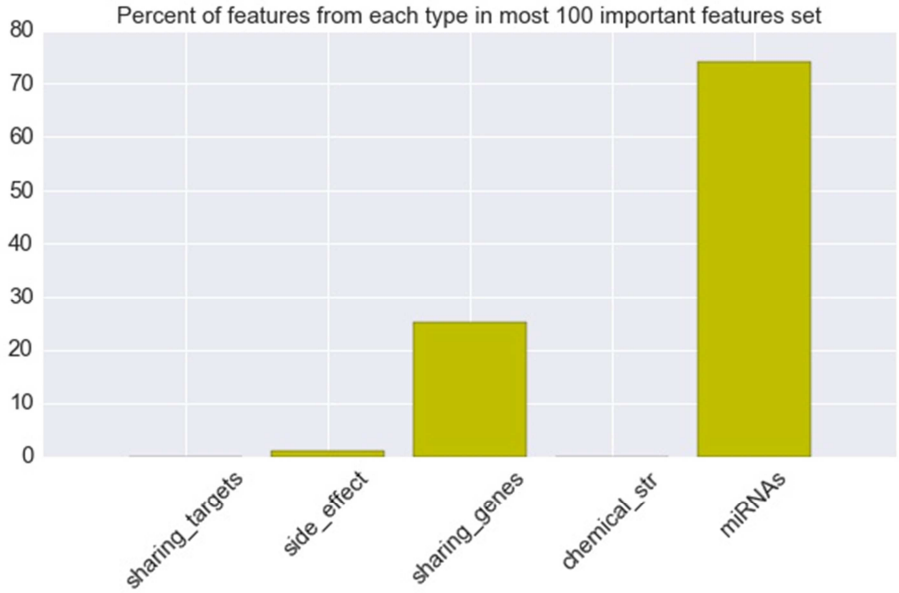


Figure 4.5 Chart for the 100 features that selected using Extra trees classifier for two classes SVM approach

B) 20 features approach: Table 4.9 contains the selected 20 features according to this method for each feature type. Figure 4.6 shows the chart of the selected 20 features, where we can see that the biggest number of important features is from miRNA type, again. Where there is no any chemical substructure, drug target or side effect features. In addition, Figure 4.7 illustrates the most weighted 20 important feature.

Table 4.9 20 features reduction using Extra trees classifier for two classes SVM approach

Feature type	Amount
Chemical substructure	0
Side effect	0
Drug target	0
Gene	6
miRNA	14

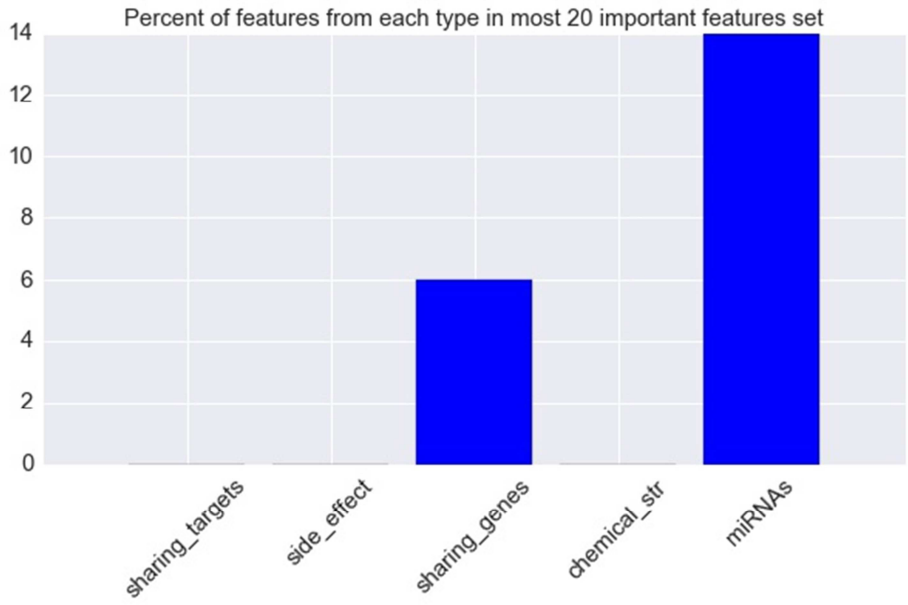


Figure 4.6 Chart for the 20 features that are selected using Extra trees classifier for two classes SVM approach

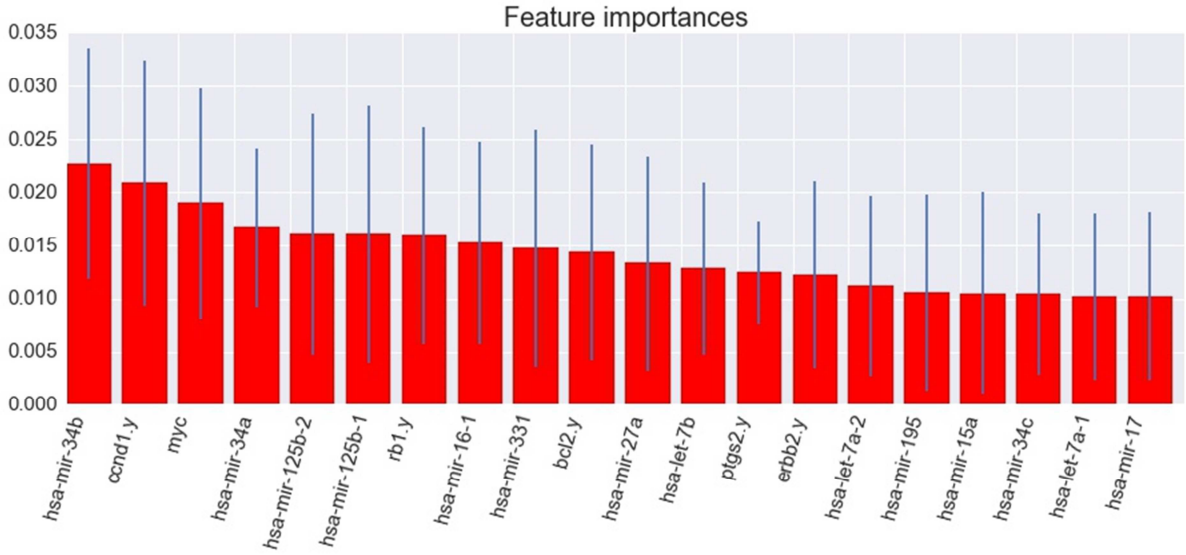


Figure 4.7 Chart for the most weighted 20 important feature according to Extra trees classifier

ii) Feature reduction using recursive feature elimination (RFE):

A) 100 features approach: Table 4.10 contains the selected 100 features according to this method for each feature type. Figure 4.8 shows the chart of the selected 100 features, where we can see that the biggest number of important features is still from miRNA type, but we can see based on this method features from side effect and chemical substructure feature types also.

Table 4.10 100 features reduction using RFE for two classes SVM approach

Feature type	Amount
Chemical substructure	11
Side effect	14
Drug target	10
Gene	23
miRNA	42

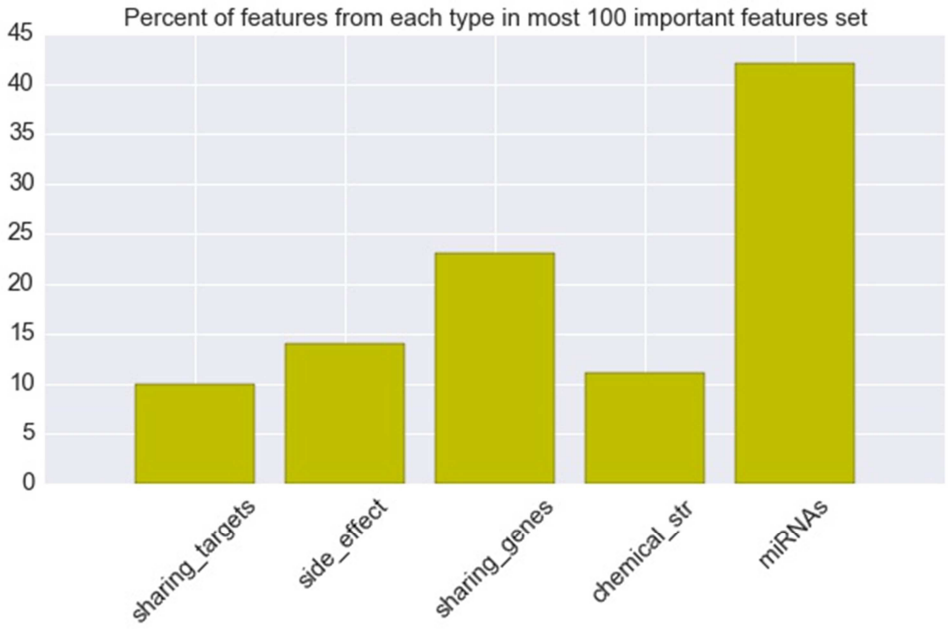


Figure 4.8 Chart for the 100 features that selected using RFE for two classes SVM approach

B) 20 features approach: Table 4.11 contains the selected 20 features according to this method for each feature type. Figure 4.9 shows the chart of the selected 20 features, where we can see that the biggest number of important features is still from miRNA and gene types, but we can see proportionally, based on this method features from side effect and chemical substructure feature types also.

Table 4.11 20 features reduction using RFE for two classes SVM approach

Feature type	Amount
Chemical substructure	1
Side effect	1
Drug target	4
Gene	6
miRNA	8

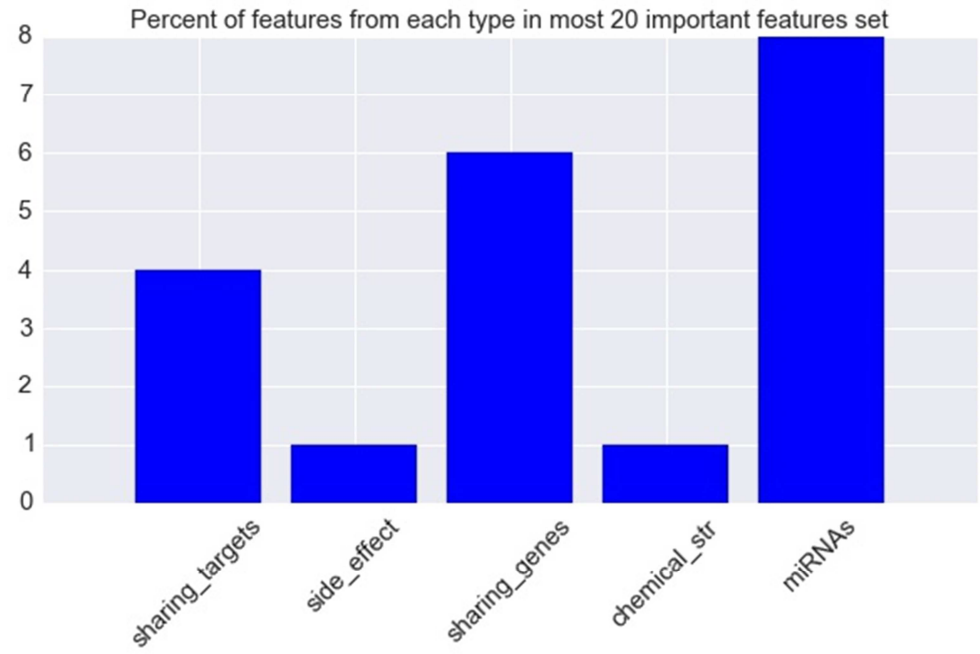


Figure 4.9 Chart for the 20 features that are selected using RFE for two classes SVM approach

iii) Feature reduction using PCA:

I also applied this principal component analysis (PCA) method, to do two approaches to reduce the features into 100 and 20. Nevertheless, due to its transformation the data set into different format by combining many features, I can not figure out the amount of the features each type based on because I can not know what is these features. But the efficiency of the method will appear after applying the model constructed using SVM.

2. Dimensionality reduction for one class classification approach: I used two methods for dimensionality reduction for one class SVM. Extra trees classifier and recursive feature elimination (RFE) methods can not be used for one class classification. Then I used state-space search method and principal component analysis PCA. Using State space search method, I was able to reduce the features into 20 features:

Table 4.12 contains the selected 20 features according to this method for each feature type. Figure 4.10 shows the chart of the selected 20 features, where we can see that the biggest number of important features is still from Chemical substructure and side effect types, and less amount of drug target and miRNA features, where the gene feature type is zero.

Table 4.12 20 features reduction using state-space search method for one class SVM approach

Feature type	Amount
Chemical substructure	10
Side effect	5
Drug target	3
Gene	0
miRNA	2

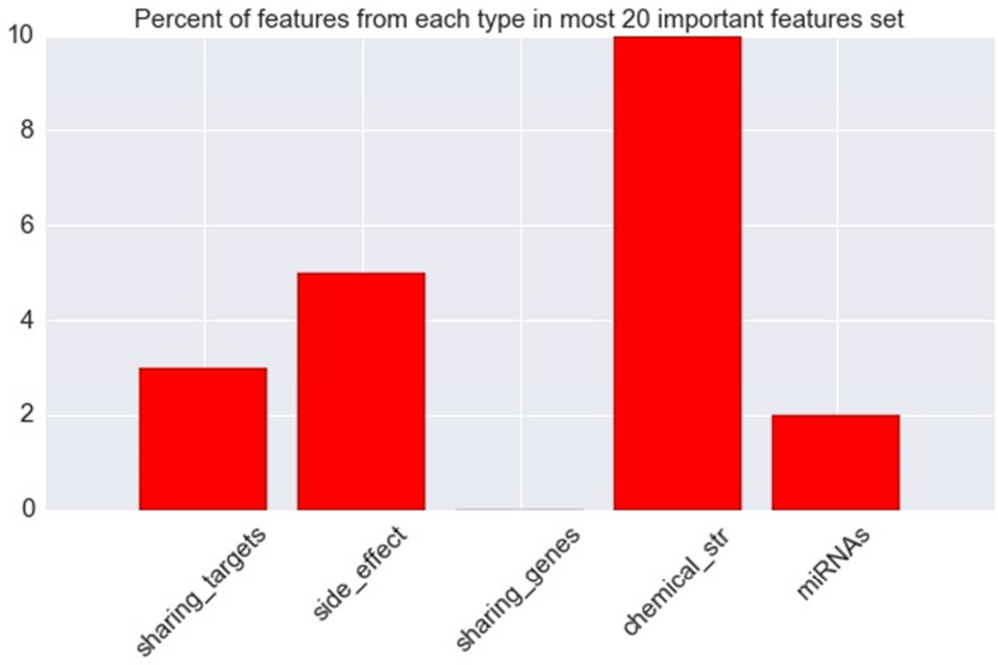


Figure 4.10 Chart for the 20 features that selected using state-space search method for one class SVM approach

4.4 **Pearson Correlation Coefficient**

The Pearson's correlation coefficient is calculated by dividing the covariance of the two variables by the product of their standard deviations. It ranges from 1 for perfectly correlated variables to -1 for perfectly anti-correlated variables and 0 means uncorrelated. Figure 4.11 shows the high and low correlations between features and classes, while Figure 4.12 shows hot plot for Pearson correlation between the important features.

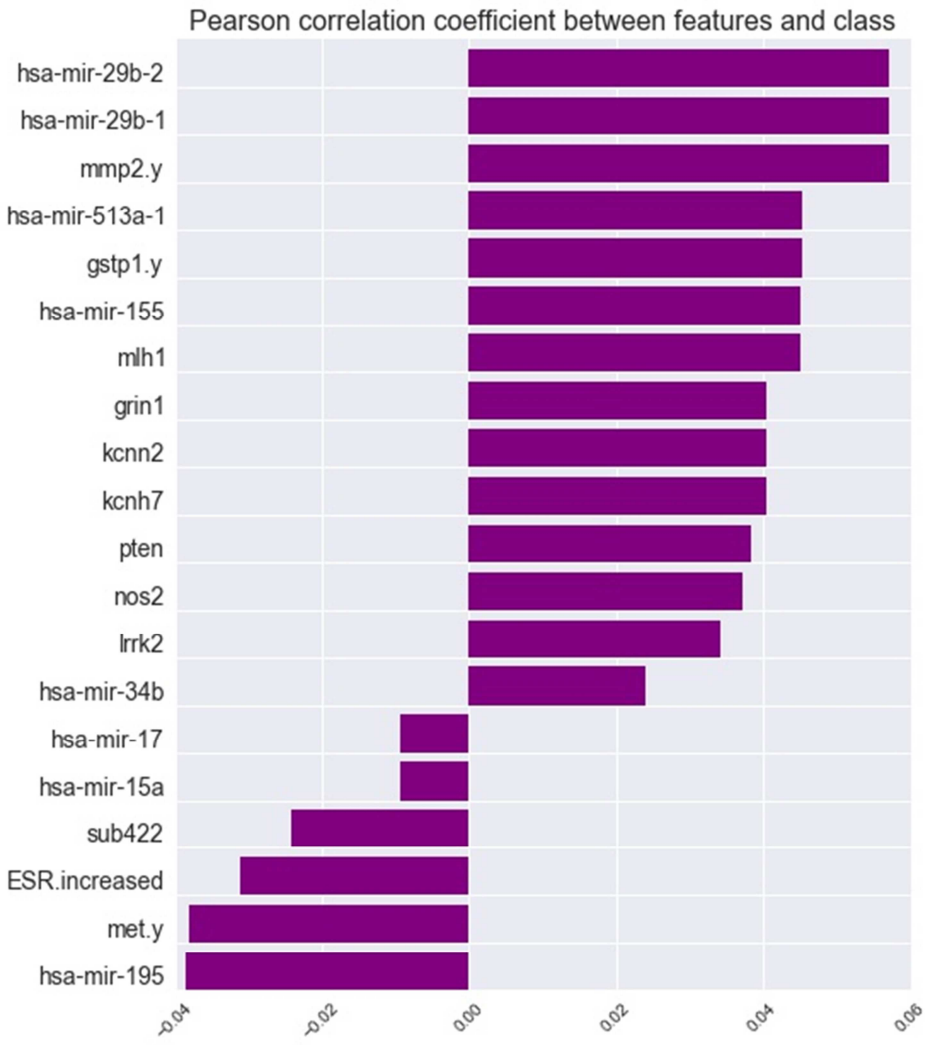


Figure 4.11 Pearson correlation coefficient between highest and lowest features and classes

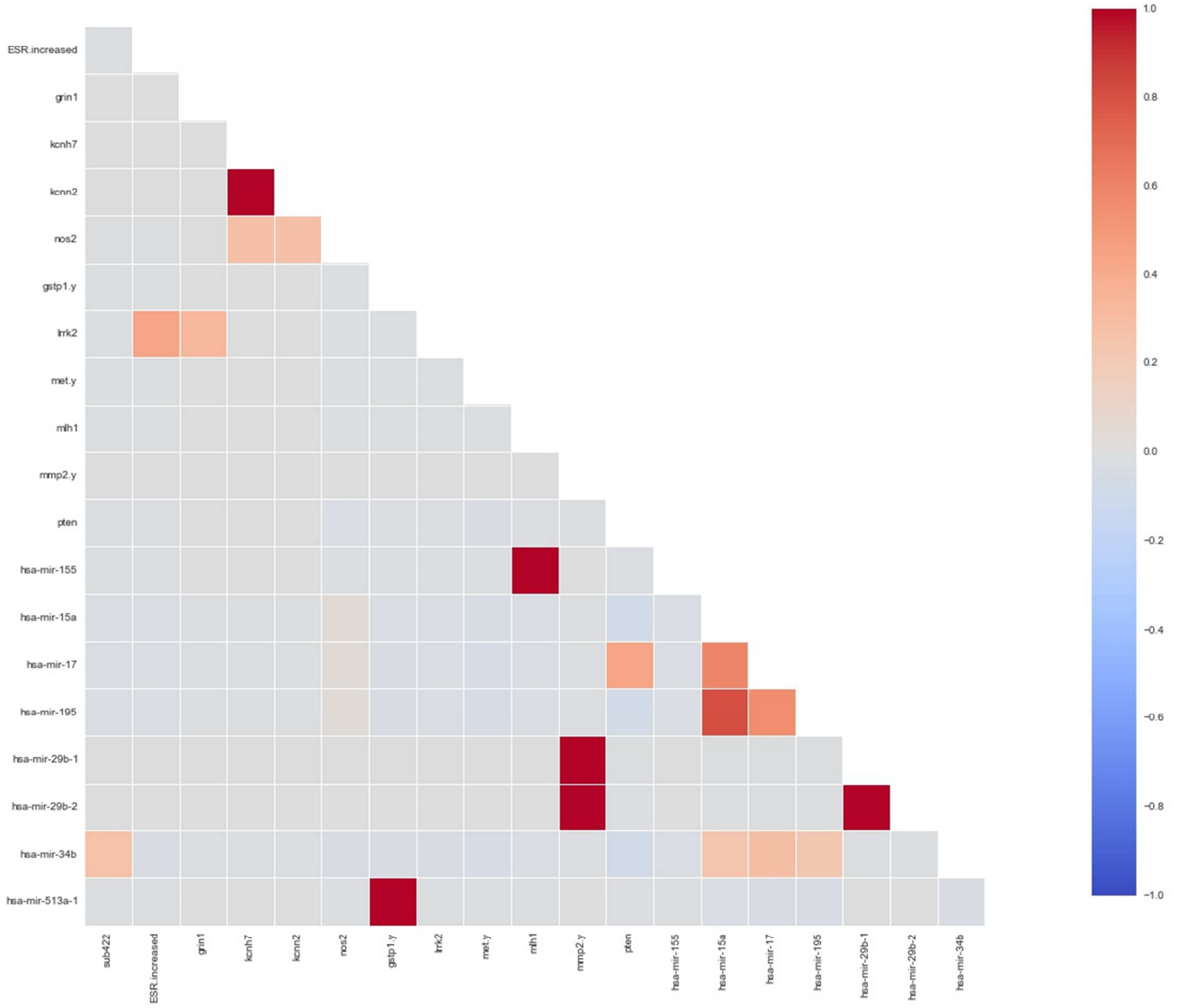


Figure 4.12 Pearson correlation coefficient between most important features

4.5 **Classification Test Accuracy**

A perfect model of classification would have an accuracy 1.00 resulting from the correct distinguish of all true positives and true negatives in the two classes' classification task and to figure out the true positives and true outliers in the one class classification. I built models and tested its accuracies after using previous reduction methods. Table 4.13 shows the results for two class classification accuracies, and Table 4.14 for one class approach.

Table 4.13 classification test accuracy for two classes approach

Method	Features number	Test set accuracy
Data set after using Extra trees classifier for reduction	100	0.265
Data set after using Extra trees classifier for reduction	20	0.00
Data set after using RFE for reduction	100	0.586
Data set after using RFE for reduction	20	0.17
Data set after using PCA for reduction	100	0.355
Data set after using PCA for reduction	20	0.33

Table 4.14 classification test accuracy for one class approach

Method	Features number	Test set accuracy
Data set after using PCA for reduction	100	0.965
Data set after using PCA for reduction	20	0.975
Data set after using state-space search for reduction	20	1.00

Classification accuracies that were obtained using PCA and state-space search methods of one class classification are extremely higher than those obtained using all the dimensionality reduction methods of two classes' classification. This shows that the randomization of drug-disease pairs that were picked up from data sets to utilized the two classes' classification was wrong and imprecise.

4.6 **The Most Predictive Features**

Since I utilized 1.00 classification test set accuracy based on dimensionality reduction using state-space search method (reduced into 20 features) and one class support vector machine (SVM) for model building, then I checked how much these features have associations either with drugs or with diseases. Table 4.15 summarizes these features, their type and their associations.

Table 4.15 Summary of most predictive features

Feature	**Type**	**Associations**	**Drug/disease**
sub571	Chemical substructure	34	Drug
sub700	Chemical substructure	17	Drug
sub413	Chemical substructure	5	Drug
sub635	Chemical substructure	34	Drug
sub338	Chemical substructure	12	Drug
sub685	Chemical substructure	24	Drug
sub393	Chemical substructure	23	Drug
sub424	Chemical substructure	3	Drug
sub359	Chemical substructure	14	Drug
sub443	Chemical substructure	18	Drug
Cancer	Side effect	17	Drug
adenomas benign	Side effect	9	Drug
arteriosclerosis	Side effect	9	Drug
leukocytosis	Side effect	3	Drug
sinus congestion	Side effect	9	Drug
tubb6	Target	19	Drug
tubb1	Target	19	Drug
kcnmb4	Target	2	Drug
microRNA 138-1	miRNA	7	Disease
microRNA 125b-2	miRNA	10	Disease

4.7 **Conclusion**

In this chapter, I prepared four hybridized data sets using drug-disease known associations and five types of features which describe these associations. In addition, I carried out three methods for dimensionality reduction for two classes classification approach, and two methods for dimensionality reduction for one class classification approach. I also generated classification models using those features which were found based on these methods.

Once a model with high accuracy is generated, in my case (state-space search for feature reduction and one class support vector machine algorithms for model generation), it can be used to predict the class of a newly drug-disease associations, and further computational and experimental analysis can be carried out in a more informed manner.

REFERENCES

[1] Sleigh, S. H., & Barton, C. L. (2010). Repurposing strategies for therapeutics. Pharmaceutical Medicine, 24(3), 151-159.

[2] Insa, R. (2013). Drug Repositioning: Bringing New Life to Shelved Assets and Existing Drugs. Edited by Michael J. Barratt and Donald E. Frail.

[3] Andronis, C., Sharma, A., Virvilis, V., Deftereos, S., & Persidis, A. (2011). Literature mining, ontologies and information visualization for drug repurposing. Briefings in bioinformatics, bbr005.

[4] Pujol, A., Mosca, R., Farrés, J., & Aloy, P. (2010). Unveiling the role of network and systems biology in drug discovery. Trends in pharmacological sciences,31(3), 115-123.

[5] Hurle, M. R., Yang, L., Xie, Q., Rajpal, D. K., Sanseau, P., & Agarwal, P. (2013). Computational drug repositioning: from data to therapeutics. Clinical Pharmacology & Therapeutics, 93(4), 335-341.

[6] O'Connor, K. A., & Roth, B. L. (2005). Finding new tricks for old drugs: an efficient route for public-sector drug discovery. Nature reviews Drug discovery,4(12), 1005-1014.

[7] Collins, F. S. (2010). Opportunities for research and NIH. Science, 327(5961), 36-37.

[8] Oprea, T. I., Bauman, J. E., Bologa, C. G., Buranda, T., Chigaev, A., Edwards, B. S., ... & Sklar, L. A. (2012). Drug repurposing from an academic perspective. Drug Discovery Today: Therapeutic Strategies, 8(3), 61-69.

[9] Lee, H. S., Bae, T., Lee, J. H., Kim, D. G., Oh, Y. S., Jang, Y., ... & Kim, S. (2012). Rational drug repositioning guided by an integrated pharmacological network of protein, disease and drug. BMC systems biology, 6(1), 80.

[10] DiMasi, J. A., Hansen, R. W., & Grabowski, H. G. (2003). The price of innovation: new estimates of drug development costs. Journal of health economics, 22(2), 151-185.

[11] Ashburn, T. T., & Thor, K. B. (2004). Drug repositioning: identifying and developing new uses for existing drugs. Nature reviews Drug discovery, 3(8), 673-683.

[12] DiMasi, J. A., Hansen, R. W., Grabowski, H. G., & Lasagna, L. (1991). Cost of innovation in the pharmaceutical industry. Journal of health economics, 10(2), 107-142.

[13] Munos, B. (2009). Lessons from 60 years of pharmaceutical innovation. Nature Reviews Drug Discovery, 8(12), 959-968.

[14] Booth, B., & Zemmel, R. (2004). Prospects for productivity. Nature Reviews Drug Discovery, 3(5), 451-456.

[15] Wu, C., Gudivada, R. C., Aronow, B. J., & Jegga, A. G. (2013). Computational drug repositioning through heterogeneous network clustering. BMC systems biology, 7(Suppl 5), S6.

[16] Xu, K., & Coté, T. R. (2011). Database identifies FDA-approved drugs with potential to be repurposed for treatment of orphan diseases. Briefings in bioinformatics, bbr006.

[17] New-York-Times-Archives (March 1998). U.s. approves sale of impo- tence pill; huge market seen. http://www.nytimes.com/1998/03/28/ us/us-approves-sale-of-impotence-pill-huge-market-seen.html? pagewanted=all&src=pm.

[18] Ghofrani, H. A., Osterloh, I. H., & Grimminger, F. (2006). Sildenafil: from angina to erectile dysfunction to pulmonary hypertension and beyond. Nature Reviews Drug Discovery, 5(8), 689-702.

[19] Jordan, V. C., Phelps, E., & Lindgren, J. U. (1987). Effects of anti-estrogens on bone in castrated and intact female rats. Breast cancer research and treatment,10(1), 31-35.

[20] FDA (September 2007). Raloxifene hydrochloride. http://www.fda.gov/ AboutFDA/CentersOffices/OfficeofMedicalProductsandTobacco/CDER/ucm129243.htm.

[21] Stephens, T., & Brynner, R. (2009). Dark remedy: the impact of thalidomide and its revival as a vital medicine. Basic Books.

[22] Barratt, M. J., & Frail, D. E. (2012). Drug repositioning: Bringing new life to shelved assets and existing drugs. John Wiley & Sons.

[23] Voelker, R. (1998). International group seeks to dispel incontinence taboo. JAMA, 280(11), 951-953.

[24] Druker, B. J. (2004). Imatinib as a paradigm of targeted therapies. Advances in cancer research, 91, 1-30.

[25] Weber, A., Casini, A., Heine, A., Kuhn, D., Supuran, C. T., Scozzafava, A., & Klebe, G. (2004). Unexpected nanomolar inhibition of carbonic anhydrase by COX-2-selective celecoxib: new pharmacological opportunities due to related binding site recognition. Journal of medicinal chemistry, 47(3), 550-557.

[26] Chow, W. A., Jiang, C., & Guan, M. (2009). Anti-HIV drugs for cancer therapeutics: back to the future? The lancet oncology, 10(1), 61-71.

[27] Bradley, D. (2005). Why big pharma needs to learn the three'R's. Nature Reviews Drug Discovery, 4(6), 446-446.

[28] Delbaldo, C., Faivre, S., Dreyer, C., & Raymond, E. (2011). Sunitinib in advanced pancreatic neuroendocrine tumors: latest evidence and clinical potential. Therapeutic advances in medical oncology, 1758834011428147.

[29] Li, Y. Y., & Jones, S. J. (2012). Drug repositioning for personalized medicine. Genome Med, 4(3), 27.

[30] Rose, J. S., & Bekaii-Saab, T. S. (2011). New developments in the treatment of metastatic gastric cancer: focus on trastuzumab. OncoTargets and therapy, 4, 21.

[31] Sekhon, B. S. (2013). Repositioning drugs and biologics: Retargeting old / existing drugs for potential new therapeutic applications. J. Pharm. Educ. Res, 4, 1-15.

[32] Fakhraei, S., Raschid, L., & Getoor, L. (2013, August). Drug-target interaction prediction for drug repurposing with probabilistic similarity logic. In Proceedings of the 12th International Workshop on Data Mining in Bioinformatics (pp. 10-17). ACM.

[33] Jin, G., & Wong, S. T. (2014). Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines. Drug discovery today,19(5), 637-644.

[34] Pfister, D. G. (2012). Off-label use of oncology drugs: the need for more data and then some. Journal of Clinical Oncology, 30(6), 584-586.

[35] Burstein, H. J. (2013). Off-label use of oncology drugs: too much, too little, or just right? Journal of the National Comprehensive Cancer Network, 11(5), 505-506.

[36] Swamidass, S. J. (2011). Mining small-molecule screens to repurpose drugs. Briefings in bioinformatics, 12(4), 327-335.

[37] Kolb, P., Ferreira, R. S., Irwin, J. J., & Shoichet, B. K. (2009). Docking and chemoinformatic screens for new ligands and targets. Current opinion in biotechnology, 20(4), 429-436.

[38] Blatt, J., & Corey, S. J. (2013). Drug repurposing in pediatrics and pediatric hematology oncology. Drug discovery today, 18(1), 4-10.

[39] McCabe, B., Liberante, F., & Mills, K. I. (2015). Repurposing medicinal compounds for blood cancer treatment. Annals of hematology, 94(8), 1267-1276.

[40] Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., ... & Golub, T. R. (2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. science, 313(5795), 1929-1935.

[41] Qu, X. A., & Rajpal, D. K. (2012). Applications of Connectivity Map in drug discovery and development. Drug discovery today, 17(23), 1289-1298.

[42] Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., ... & Myer, V. E. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature, 483(7391), 603-607.

[43] Sirota, M., Dudley, J. T., Kim, J., Chiang, A. P., Morgan, A. A., Sweet-Cordero, A., ... & Butte, A. J. (2011). Discovery and preclinical validation of drug indications using compendia of public gene expression data. Science translational medicine, 3(96), 96ra77-96ra77.

[44] Sanseau, P., Agarwal, P., Barnes, M. R., Pastinen, T., Richards, J. B., Cardon, L. R., & Mooser, V. (2012). Use of genome-wide association studies for drug repositioning. Nature biotechnology, 30(4), 317-320.

[45] Zhang, J., Jiang, K., Lv, L., Wang, H., Shen, Z., Gao, Z., ... & Wang, S. (2015). Use of Genome-Wide Association Studies for Cancer Research and Drug Repositioning. PloS one, 10(3), e0116477.

[46] Lussier, Y. A., & Chen, J. L. (2011). The emergence of genome-based drug repositioning. Science translational medicine, 3(96), 96ps35-96ps35.

[47] Zhao, H., Jin, G., Cui, K., Ren, D., Liu, T., Chen, P., ... & Wong, S. T. (2013). Novel modeling of cancer cell signaling pathways enables systematic drug repositioning for distinct breast cancer metastases. Cancer research, 73(20), 6149-6163.

[48] Iskar, M., Zeller, G., Blattmann, P., Campillos, M., Kuhn, M., Kaminska, K. H., ... & Bork, P. (2013). Characterization of drug-induced transcriptional modules: towards drug repositioning and functional understanding. Molecular systems biology, 9(1), 662.

[49] Johnson, M. A., & Maggiora, G. M. (1990). Concepts and applications of molecular similarity. Wiley.

[50] Activity relationship https://en.wikipedia.org/wiki/Quantitative_structure%E2%80%93activity_relationship. acess time, 27.12.2015

[51] Nikolova, N., & Jaworska, J. (2003). Approaches to measure chemical similarity–a review. QSAR & Combinatorial Science, 22(9-10), 1006-1026.

[52] Noeske, T., Sasse, B. C., Stark, H., Parsons, C. G., Weil, T., & Schneider, G. (2006). Predicting Compound Selectivity by Self-Organizing Maps: Cross-Activities of Metabotropic Glutamate Receptor Antagonists. ChemMedChem,1(10), 1066-1068.

[53] Keiser, M. J., Setola, V., Irwin, J. J., Laggner, C., Abbas, A. I., Hufeisen, S. J., ... & Roth, B. L. (2009). Predicting new molecular targets for known drugs. Nature, 462(7270), 175-181.

[54] Croset, S. (2014). Drug repositioning and indication discovery using description logics (Doctoral dissertation, University of Cambridge).

[55] Iorio, F., Bosotti, R., Scacheri, E., Belcastro, V., Mithbaokar, P., Ferriero, R., ... & di Bernardo, D. (2010). Discovery of drug mode of action and drug repositioning from transcriptional responses. Proceedings of the National Academy of Sciences, 107(33), 14621-14626.

[56] Meng, X. Y., Zhang, H. X., Mezei, M., & Cui, M. (2011). Molecular docking: a powerful approach for structure-based drug discovery. Current computer-aided drug design, 7(2), 146.

[57] Haupt, V. J., & Schroeder, M. (2011). Old friends in new guise: repositioning of known drugs with structural bioinformatics. Briefings in bioinformatics, bbr011.

[58] Kuhn, M., Campillos, M., Letunic, I., Jensen, L. J., & Bork, P. (2010). A side effect resource to capture phenotypic effects of drugs. Molecular systems biology, 6(1), 343.

[59] Witten, I. H., Don, K. J., Dewsnip, M., & Tablan, V. (2004). Text mining in a digital library. International Journal on Digital Libraries, 4(1), 56-59.

[60] Barçante1i, E., Jezuz1i, M., Duval, F., Caffarena, E., Cruz, O. G., & Silva, F. Identifying Drug Repositioning Targets Using Text Mining.

[61] Patchala, J., & Jegga, A. G. (2015). Concept Modeling-based Drug Repositioning. AMIA Summits on Translational Science Proceedings, 2015, 222.

[62] Machine learning definition, http://ai.stanford.edu/~ronnyk/glossary.html. Acess time. 27.12.2015.

[63] Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., ... & Robles, V. (2006). Machine learning in bioinformatics. Briefings in bioinformatics, 7(1), 86-112.

[64] Napolitano, F., Zhao, Y., Moreira, V. M., Tagliaferri, R., Kere, J., D'Amato, M., & Greco, D. (2013). Drug repositioning: a machine-learning approach through data integration. J. Cheminformatics, 5, 30.

[65] Gottlieb, A., Stein, G. Y., Ruppin, E., & Sharan, R. (2011). PREDICT: a method for inferring novel drug indications with application to personalized medicine. Molecular systems biology, 7(1), 496.

[66] Von Eichborn, J., Murgueitio, M. S., Dunkel, M., Koerner, S., Bourne, P. E., & Preissner, R. (2011). PROMISCUOUS: a database for network-based drug-repositioning. Nucleic acids research, 39(suppl 1), D1060-D1066.

[67] Ewing, T. J., Makino, S., Skillman, A. G., & Kuntz, I. D. (2001). DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. Journal of computer-aided molecular design, 15(5), 411-428.

[68] Luo, H., Chen, J., Shi, L., Mikailov, M., Zhu, H., Wang, K., ... & Yang, L. (2011). DRAR-CPI: a server for identifying drug repositioning potential and adverse drug reactions via the chemical–protein interactome. Nucleic acids research, gkr299.

[69] Liu, C. C., Tseng, Y. T., Li, W., Wu, C. Y., Mayzus, I., Rzhetsky, A., ... & Zhou, X. J. (2014). DiseaseConnect: a comprehensive web server for mechanism-based disease–disease connections. Nucleic acids research, 42(W1), W137-W146.

[70] Abu-Mostafa, Y. S., Magdon-Ismail, M., & Lin, H. T. (2012). Learning from data. AMLBook.

[71] Matlab documentations (2015), Matlab Statistics and Machine Learning Toolbox User's Guide.

[72] Wang, Y., Chen, S., Deng, N., & Wang, Y. (2013). Drug repositioning by kernel-based integration of molecular structure, molecular activity, and phenotype data.

[73] Yang, L., & Agarwal, P. (2011). Systematic drug repositioning based on clinical side-effects. PloS one, 6(12), e28025.

[74] Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A. C., Liu, Y., ... & Wishart, D. S. (2014). DrugBank 4.0: shedding new light on drug metabolism. Nucleic acids research, 42(D1), D1091-D1097.

[75] Drugbank statistics, http://www.drugbank.ca/stats, access time. 27.12.2015

[76] Bolton, E. E., Wang, Y., Thiessen, P. A., & Bryant, S. H. (2008). PubChem: integrated platform of small molecules and biological activities. Annual reports in computational chemistry, 4, 217-241.

[77] Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research, 28(1), 27-30.

[78] KEGG statistics, http://www.kegg.jp/kegg/docs/statistics.html, acess time, 27.12.2015

[79] Kuhn, M., Campillos, M., Letunic, I., Jensen, L. J., & Bork, P. (2010). A side effect resource to capture phenotypic effects of drugs. Molecular systems biology, 6(1), 343.

[80] Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., ... & Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic acids research, 42(D1), D1001-D1006.

[81] Kubinyi, H. (2002). Chemical similarity and biological activities. Journal of the Brazilian Chemical Society, 13(6), 717-726.

[82] Jaccard index, https://en.wikipedia.org/wiki/Jaccard_index , access time. 27.12.2015

[83] Rogers, D. J., & Tanimoto, T. T. (1960). A computer program for classifying plants. Science, 132(3434), 1115-1118.

[84] Li, J., & Lu, Z. (2015). An Integrative Approach for Discovery of New Uses of Existing Drugs. Data Science Journal, 14.

[85] Chen, H., & Zhang, Z. (2015). A miRNA-Driven Inference Model to Construct Potential Drug-Disease Associations for Drug Repositioning. BioMed research international, 2015.

[86] Dimensionality reduction, https://en.wikipedia.org/wiki/Dimensionality_reduction, access time, 27.12.2015

[87] Support vector machine, https://www.kuleuven.be/samenwerking/iminds/medicalit/examples/case-clinical-decision-support-stadius, access time, 27.12.2015

[88] Dimensionality reduction figures, http://www.glassbeam.com/scalable-machine-learning-apache-spark-mlbase/, access time, 27.12.2015

[89] Foster provost, Tom Fawcett (2013), Data science for business. O'Reilly media.

[90] Support vector machine, http://www.mathworks.com/help/stats/support-vector-machines-svm.html?refresh=true , access time, 27.12.2015

[91] Curse dimensionality, http://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/, access time, 27.12.2015

[92] Kernel methods versus traditional methods figures , http://videolectures.net/mlsb2012_moreau_kernel/, access time, 27.12.2015

[93] DRAR-CPI figure http://cpi.bio-x.cn/drar/, access time, 27.12.2015

[94] SIDER, side effect database http://sideeffects.embl.de/, access time, 26.1.2016

[95] PubChem, open chemistry database, https://pubchem.ncbi.nlm.nih.gov/, access time, 26.1.2016

[96] OMIM, online Mendelian inheritance in Man, http://www.omim.org/, access time, 26.1.2016

[97] MLP, molecular libraries program, http://mli.nih.gov/mli/, access time, 26.1.2016

[98] NCRR, http://www.ncrr.nih.gov/, access time, 26.1.2016

[99] IMI, innovative medicines initiative, http://www.imi.europa.eu/, access time, 26.1.2016

[100] GEO, gene expression omnibus, http://www.ncbi.nlm.nih.gov/geo/, access time, 26.1.2016

[101] SRA, sequence read archive, http://www.ncbi.nlm.nih.gov/sra, access time, 26.1.2016

[102] Mantra 2.0, mode of action. http://mantra.tigem.it/, access time, 26.1.2016

[103] PharmDB-K, http://www.pharmdb.org/, access time, 26.1.2016

[104] PROMISCUOUS, http://bioinformatics.charite.de/promiscuous/, access time, 26.1.2016

[105] Disease-Connect, web database for mechanism-based disease-disease connections, http://disease-connect.org/, access time, 26.1.2016

[106] DRUGBANK, drug and drug target database, http://www.drugbank.ca/, access time, 26.1.2016

[107] KEGG, Kyoto encyclopedia of genes and genomes, http://www.genome.jp/kegg/, access time, 26.1.2016

CURRICULUM VITAE

Name Surname	**:** Abdullah ALRHMOUN
Place and Date of Birth	**:** Saudi Arabia, 04/ 07/ 1987
Address	**:** Başakşehir Mah. Sancak Sk. Başakşehir 1. Etap Sitesi B25 Blok No: 26 D:1 Başakşehir / Istanbul /Turkey
E-Mail	**:** abdullahalrhmoun@gmail.com
B.Sc.	**:** Al-Baath University, Health Science Faculty, Physiotherapy (2010)