Drug repositioning and indication discovery using description logics

Samuel Claude Jean Croset

Darwin College

A thesis submitted on March, 2014 for the Degree of Doctor of Philosophy

Abstract

Drug repositioning is the discovery of new indications for approved or failed drugs. This practice is commonly done within the drug discovery process in order to adjust or expand the application line of an active molecule. Nowadays, an increasing number of computational methodologies aim at predicting repositioning opportunities in an automated fashion. Some approaches rely on the direct physical interaction between molecules and protein targets (docking) and some methods consider more abstract descriptors, such as a gene expression signature, in order to characterise the potential pharmacological action of a drug (Chapter 1).

On a fundamental level, repositioning opportunities exist because drugs perturb multiple biological entities, (on and off-targets) themselves involved in multiple biological processes. Therefore, a drug can play multiple roles or exhibit various mode of actions responsible for its pharmacology. The work done for my thesis aims at characterising these various modes and mechanisms of action for approved drugs, using a mathematical framework called description logics.

In this regard, I first specify how living organisms can be compared to complex black box machines and how this analogy can help to capture biomedical knowledge using description logics (Chapter 2). Secondly, the theory is implemented in the Functional Therapeutic Chemical Classification System (FTC - `https://www.ebi.ac.uk/chembl/ftc/`), a resource defining over 20,000 new categories representing the modes and mechanisms of action of approved drugs. The FTC also indexes over 1,000 approved drugs, which have been classified into the mode of action categories using automated reasoning. The FTC is evaluated against a gold standard, the Anatomical Therapeutic Chemical Classification System (ATC), in order to characterise its quality and content (Chapter 3).

Finally, from the information available in the FTC, a series of drug repositioning hypotheses were generated and made publicly available via a web application (`https://www.ebi.ac.uk/chembl/research/ftc-hypotheses`). A sub-

set of the hypotheses related to the cardiovascular hypertension as well as for Alzheimer's disease are further discussed in more details, as an example of an application (Chapter 4).

The work performed illustrates how new valuable biomedical knowledge can be automatically generated by integrating and leveraging the content of publicly available resources using description logics and automated reasoning. The newly created classification (FTC) is a first attempt to formally and systematically characterise the function or role of approved drugs using the concept of mode of action. The open hypotheses derived from the resource are available to the community to analyse and design further experiments.

Declaration

This thesis:

- is my own work and contains nothing which is the outcome of work done in collaboration with others, except where specified in the text;
- is not substantially the same as any that I have submitted for a degree or diploma or other qualification at any other university; and
- does not exceed the prescribed limit of 60,000 words (approximately 39,853 words).

Samuel Claude Jean Croset
March, 2014

Acknowledgements

Writing a thesis is a challenging enterprise, and I would like to thank all the people who provided support along the journey. First to both my supervisors, Dietrich Rebholz-Schuhmann (DRS) and John Overington (JPO) for giving me the opportunity and freedom of doing my own science. Then to all the folks who provided me valuable advice and enlightening discussion, in particular Sarah Carl, Robert Hoehndorf, Anika Oellrich, Felix Krueger, Rita Santos, Christoph Grabmuller, Benjamin Stauch, John May, Gerard van Westen, George Papadatos, Grace Mugambate, Syed Asad Rahman and Pedro Ballester. I have interacted mostly with two research groups during my PhD time at the EMBL-EBI, the text-mining and the ChEMBL group; I would like to show my gratitude to all the members of both of these teams. I would also like to include my family in the acknowledgements, in particular my mother Evelyne, my father Serge and my little brother Elliott, for providing mental support and cheese. All my friends should be present in this list, both from here such as the predocs or out from of the country such as the Haute-Savoie folks. A special thanks to Sarah Carl (<3) and Jean-Baptiste Pettit, as well as to Jacek for the spiritual support. Finally I would like to sincerely thank the reader, for taking the time and energy to go through this document. I deeply apologize in advance for the numerous people I did not mention here, and I will offer a free pub lunch to thank anyone reading this manuscript - just send me an email (samuel.croset@gmail.com) with the subject "THESIS ACKNOWLEDGEMENT" to claim your meal.

Contents

chapter 1

Review of computational drug repositioning approaches

Key points

- Drug repositioning is the discovery of new indications for approved or failed drugs.
- Repositioning opportunities exist because drugs perturb multiple biological entities (on and off-targets) themselves involved in multiple biological processes. As the drug discovery pipelines are focused on one disease of interest, a therapeutic application for a drug to other areas can be missed.
- A variety of predictive computational approaches have been developed to identify drug repositioning opportunities, each based around a biological concept of interest, or abstraction process over the data.
- For my thesis, I decided to investigate drug repositioning using the concept of *mode of action*. This notion is of interest as it can represent the various biological functions a drug can play in an organism and help to identify new therapeutic applications.

Author's comment

This chapter is a review of the computational methods developed in the last decade, presenting their relative strengths and weaknesses. This state-of-the-art analysis helped me to define the rationale for my thesis.

Living organisms, just as machines, are subject to dysfunction. Sometimes an internal abnormality can impair the correct functioning or sometimes external forces, such as the interaction with the environment, can damage a body. If not fixed, malfunctions accumulate and will eventually result in the death of the entity, namely the cessation of all functions. Since prehistoric times (Wikipedia, 2014g), humans have been interested in preventing and handling dysfunctions, in order to extend the lifespan of objects or to improve the quality of their own existence. When the aim is to fix living bodies, this practice goes by the name of *medicine*, or the art of preventing, diagnosing and treating diseases.

Alleviating an impairment is a daunting task, from both a biological and legal perspective. In this regard, doctors can rely on a number of tools, engineered throughout the years. Among the most commonly used ones such as medical devices, are the active small molecules, the drugs. Drugs are of primary interest, as they can impact the treatment of complex processes with molecular roots, such as cancer or pain for instance, in a relatively controllable and safe manner. However, in order to be usable in clinics, a candidate drug has first to go through a development phase which takes nowadays at least a dozen years and can cost up to a billion dollars (DiMasi, 2001). This process involves a myriad of different people, from biologists to law attorneys.

As my interests revolve primarily around pharmacology and computer sciences, I decided to focus my efforts on the molecular side of the problem. More precisely, I decided to systematically characterise and understand the multiple roles any drug can play in the human body, using computational means. My work will be extensively described throughout this manuscript and finds an application in a topic named *drug repositioning*, which is the object of this chapter.

1.1 Relevance of drug repositioning

Drug repositioning (also referred as *drug repurposing*, *re-profiling*, *therapeutic switching* and *drug re-tasking*) is *the identification of new therapeutic indications for known drugs*. These drugs can either be approved and marketed compounds used daily in a clinical setting, or they can be drugs that have been "shelved", namely molecules that did not succeed in clinical trials or for which projects have

been discontinued for various reasons. In one sentence, drug repositioning can be defined as *renewing failed drugs and expanding successful ones* (Barratt and Frail, 2012).

One motivation behind drug repositioning is the possibility to further market and extend the application line or patent life of a drug, therefore increasing the revenue stream generated from it. Another aim is the treatment of rare or neglected diseases; usually such conditions are difficult to address for financial reasons, yet there might exist some safe and active molecules already developed for other indications, deemed suitable for this scenario (Men et al., 2010). I refer the reader to some recent excellent reviews (Ashburn and Thor, 2004) (Dudley et al., 2011a) (Hurle et al., 2013) in order to fully appreciate the economical market behind this approach, as well as legal challenges coming along. I limited myself, in the context of this work, to exploring the subject from an academic perspective: characterising the various roles approved drugs can play using computational means, without necessarily seeking business opportunities.

1.1.1 Opportunities for finding new indications

Many drugs have been successfully repositioned in the past; classical examples such as sildenafil (Viagra) and thalidomide will be presented in the coming sections. But first, the scientific legitimacy of the idea behind drug repurposing should be discussed: how is it possible for a drug to play multiple roles? What is the molecular rationale? Why does the drug discovery process not automatically identify such opportunities?

In order to be able to answer these questions, I will briefly present the traditional drug development pipeline, which produced most of the recent therapeutic chemicals (Swinney and Anthony, 2011).

The fundamental idea behind the discovery of a new medicine has not evolved much since prehistoric times (Wikipedia, 2014g); it still consists mostly of a trial and error process, hence the term *discovery*. The operation starts by picking a disease of interest, selected in terms of market size or clinical needs. Then a large collection of chemicals is experimentally tried to see if any relevant effect in regards to the chosen condition can be produced. This procedure is called *screening*. There exist mostly two types of screening: target-based and phenotypic. The former optimises the selection of the chemicals on the ability to bind

a biological entity (usually a protein), called the target and that is relevant for the pathological process studied. The more specifically the chemical interacts with the active site of the target, the cleaner the action of the chemical will be (*key-lock* model - see Figure 1.1).

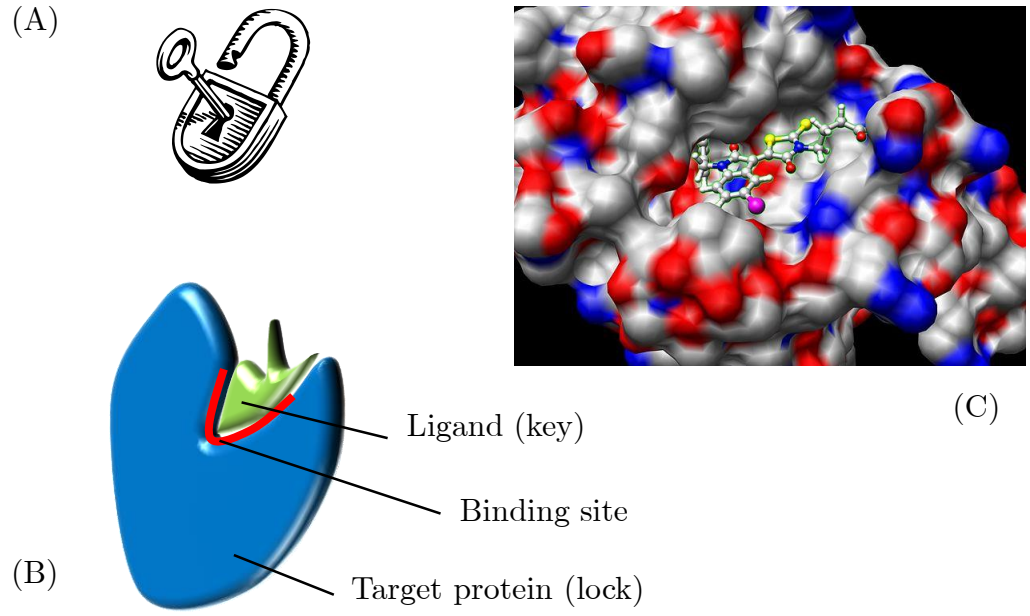


Figure 1.1: Key-lock model. (A) Illustration of the analogy. (B) Schematic representation of the interaction between a ligand and a protein (receptor, enzyme, transporter, etc.). One aim of drug discovery is to find alternative molecules (i.e. drugs) capable of either mimicking the action of the natural ligand or reversing it. In theory, the more specific the molecule is to the binding site, the more accurate the pharmacological response will be. (C) Computer generated representation of the interaction between a ligand and a protein. (Source `http://en.wikipedia.org/wiki/File:Docking.jpg`)

On the other hand, a phenotypic screen does not make any assumptions about the underlying pathological mechanism and proteins involved. A cell line or model organism representative of the disease is directly used to read the results of the screening. Both methods help to efficiently discover active chemicals, whose structures are then further optimised for efficacy (so called *lead optimisation*). Now drug repurposing opportunities can be derived from three key observations (Barratt and Frail, 2012) about the traditional workflow presented above and summarised in Figure 1.2.

The first observation appreciates the limits of the key-lock model and the "magic bullet" concept (Wikipedia, 2014e). In practice, it is challenging to design

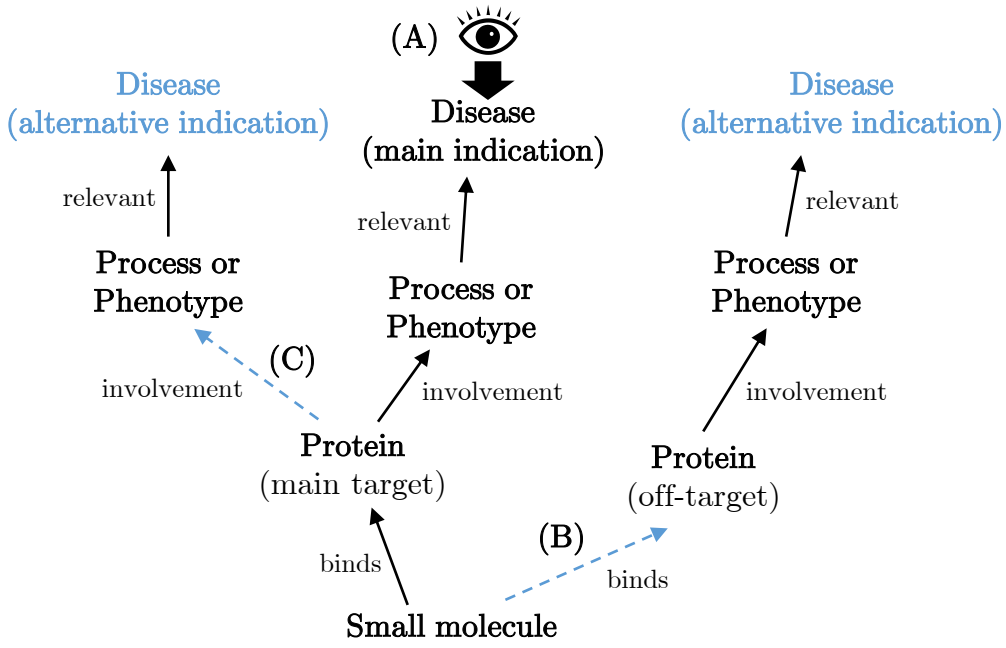


Figure 1.2: Drug discovery process and opportunities for drug repositioning (in blue; logical connection at the origin of the opportunity are dashed). (A) Traditional drug discovery workflow: from biological evidence showing the involvement of a protein in a process relevant to a particular disease of interest, a series of chemicals are screened for bioactivity. Potent and safe molecules will become drugs indicated for the disease. Paul Ehrlich postulated that chemicals should be as specific as possible against the target (organism or protein), in order to increase the control over the pharmacology. The concept of "magic bullet" describes such an ideal therapeutic compound. (B) Repositioning hypothesis: often, a small molecule binds to other proteins (so called *off-targets*), themselves potentially involved in other pathologies. (C) Repositioning hypothesis: the main protein target can be involved in a series of alternative biological processes and relevant for another disease.

a molecule that will interact with a single protein only (Paolini et al., 2006) (Li et al., 2010). Most of the drugs will bind to multiple proteins in an organism and can therefore produce a variety of unwanted effects (concerns between 35 to 62% of the chemicals - sometimes described as "drug promiscuity"). This characteristic is well-known and referred to as *off-target* effect or *polypharmacology*. Identifying the off-target proteins interacting with known drugs can provide repurposing opportunities.

The second observation values evidence that molecular targets are themselves involved in multiple biological processes and capable of performing multiple functions. Historically speaking, one protein target was assumed to be responsible

for one biological role. Hitting the right protein, such as a receptor or an enzyme known to be directly involved in the disease was a straightforward way to generate relevant phenotypic outcomes, as illustrated in Figure 1.2. However, this approach overlooks the fact that proteins function in the context of signalling pathways and networks (Hopkins, 2008) (Barabasi and Oltvai, 2004), the perturbations experienced by one node cascading on to its neighbours due to molecular dynamism. Therefore understanding and appreciating the involvement of the protein target in the biological system can help to identify a new role for a drug.

The last observation derives from the very strategy used by drug discovery programs. As a disease or condition is the starting point of the screening, identified chemicals are by definition optimised in this regard. Compounds will be tested during the expensive clinical trials for the main indication mostly and obviously not for all possible diseases. It is therefore possible for a compound to be used for different purposes, yet not be identified as such. Related to this, new usages can be discovered during clinical practice, and the prescription of the drug can evolve accordingly even if the alternative indication is not legally approved. This practice is known as off-label prescription (Wikipedia, 2014c) and demonstrates the potential of a drug to address various indications.

To conclude, these three observations alone justify the potential existence of repositioning opportunities, especially for the drugs discovered via the target-based or phenotypic screening. Meaningful repurposing options appear to be strongly dependent on the available biomedical knowledge, as well as our understanding of the molecular system.

1.1.2 Drug repositioning faces legal and scientific challenges

In theory, it is possible for a chemical to be active for multiple therapeutic indications. Yet in practice, several obstacles can impair the development of a potential new usage.

The first factor to handle is serendipity. Biology is complex and capricious, exemplified by diseases such as cancers or dementias (Ashburn and Thor, 2004). A drug rarely keeps its original indication (Barratt and Frail, 2012); it gets reoriented throughout the years when more data becomes available and its *in vivo* pharmacology better understood. Famous repurposing and discovery stories were

mostly due to chance and unexpected results (e.g. the Viagra, as will be discussed), therefore it can be difficult to forecast any relevant opportunities.

Other challenges come from the corporate and legal aspect around drug discovery. Indeed, in order to be commercially valid, a molecule needs much more than just to exhibit powerful pharmacological features. If the intellectual properties around the molecule are expired or close to be, there might be no incentive to continue the research on an alternative indication, as no profit would derive from it. Re-indicating a drug is also not part of the standard regulatory procedures, therefore administrative problems can happen, delaying or preventing the new usage of the compound. Moreover, some unnecessary safety concerns could appear when the drug is tested for the new indication. Indeed, the repositioning would potentially target a different group of patients, with different physiological conditions, and it is not to be excluded that an unforeseen adverse event could happen during the trials over the new population, hence compromising the original indication.

The dosage at which the drug is administered could also be a potential obstacle; the molecule still needs to preserve a good efficacy and show some activity at low concentration. Depending on the anatomical part of the body targeted, the formulation should also be reconsidered for efficacy. These factors can alter the pharmacokinetic profile of the drug and compromise the safety of the patient.

To conclude, the molecular opportunities for drug repositioning are contrasted by practical challenges. Even if a compound is found to be active and safe for a new indication, additional factors, in particular legal issues and intellectual property, have to be considered in order to successfully bring the molecule to the market.

1.2 Drug repositioning and indication discovery: success stories

I discussed briefly in the previous section the theoretical legitimacy of drug repositioning and its potential limitations. I will now present three success stories, exemplifying the theory and showing evidence that repositioning can happen in practice. These case-scenarios are of particular interest as they each illustrate a different reason behind the repurposing. Understanding the logic backing the

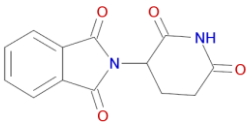
findings is paramount in order to build successful predictive methods later on.

1.2.1 Sildenafil: repositioning from clinical side-effects

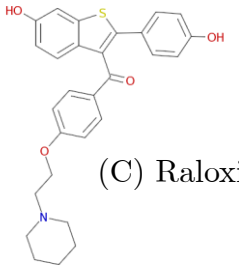
The National Health Service (NHS) defines angina as *a chest pain that occurs when the blood supply to the muscles of the heart is restricted. It usually happens because the arteries supplying the heart become hardened and narrowed* (NHS-Choices, 2014a). Sildenafil (see structure in Figure 1.3) was originally developed for this condition in the late 1980s. The working hypothesis was that an inhibition of the activity of the phosphodiesterase-5 (PDE5), an enzyme controlling the relaxation of the coronary arteries, should increase the blood flow and release the symptoms in the patient. Unfortunately, during the clinical trials, the drug lacked efficacy in regard to angina and development was discontinued, until patients shyly started to report an unusual side-effect: prolonged erections (personal discussion with molecule's investigator). Pfizer's scientists therefore decided to investigate the drug for this indication and a worldwide study of 3,700 men confirmed the effectiveness of the molecule (New-York-Times-Archives, 1998). As the pharmacokinetic profile was suitable, the drug was repurposed towards erectile dysfunction accordingly. Viagra became a blockbuster drug, with annual sales higher than 1.5 billion dollars (Renaud and Xuereb, 2002) during the first years of its release. The molecule was first-in-class for this indication and impacted the social life of millions of humans (Renaud and Xuereb, 2002). The story does not end here; the drug is currently used for pulmonary arterial hypertension too, after demonstrating a successful improvement in patients during clinical trials (Ghofrani et al., 2006).

What can be learned from this story? Firstly, the repositioning opportunities came from secondary functions of the enzyme targeted. PDE5 was known to be involved in the erection process (Krall et al., 1988), therefore the opportunity could have been logically identified rather than being discovered by pure chance. In this regard, characterising the systematic function of protein targets can therefore lead to repositioning hypotheses. Secondly, it was necessary to wait for clinical trials in order to observe the true behaviour of the drug. This observation stresses the known difficulty of moving from cell-based assays into the human body: physiology is complex and the expected outcomes are not necessarily met in practice. Lastly, the more is known about the compound's pharmacology, the

(A) Sildenafil



(B) Thalidomide



(C) Raloxifene

Figure 1.3: Examples of drug successfully repurposed. (A) Sildenafil molecular structure and picture of tablets used for administration. (B) Thalidomide molecular structure. The teratogenic effect of the drug is illustrated on the picture. (C) Molecular structure of the raloxifene. Illustration are from Wikipedia, chemical structures from the ChEMBL database.

more roles the molecule can play, as shown with the indication for the pulmonary arterial hypertension, which appeared years later. An understanding of the internal logic of biological organisms should therefore help to predict such cases.

1.2.2 Thalidomide: repositioning a hazardous drug

The story of thalidomide illustrates how a drug can surprisingly come back from being a hazardous drug retracted from the market into a novel and unique therapeutic agent. The chemical started its life in 1957 in Europe, as a sedative, sleep-inducing agent, structurally close to barbiturates (see thalidomide structure on Figure 1.3). Because of these pharmacological properties, thalidomide was marketed to treat morning sickness in pregnant women. The drug was assumed to be safe, based on an *in vivo* studies in rodents (Stephens and Brynner, 2009). Tragically, this was not the case for humans; the drug caused severe skeletal birth defects in children born from women taking the drug. Over 15,000 newborns

were affected, suffering from anatomical malformations (see Figure 1.3). Because of this disastrous side-effect, the molecule was quickly withdrawn and triggered important reforms in the drug regulatory system (Stephens and Brynner, 2009). The story could have ended here, if it were not for an incidental discovery by Jacob Sheskin. The practitioner was trying to treat patients affected by *erythema nodosum leprosum*, a particularly painful inflammatory condition characterised by red nodules under the skin. An evening of 1964, an affected patient could not sleep as the pain was so intense. Sheskin decided to ultimately use some thalidomide, as the compound was known for its potent sleep-inductive properties and was available in this hospital. The drug worked and the patient was well rested in the morning. And as a general surprise, all pain and soreness disappeared overnight too. Intrigued by the idiosyncratic effect of the drug, Sheskin further studied the action of thalidomide in clinical trials (Barratt and Frail, 2012) and successfully showed that the drug can indeed treat *erythema nodosum leprosum* in two weeks' time in most subjects. Thalidomide found a new life and became the first and only drug approved for this indication. Just as for the sildenafil molecule, the potential of the chemical is still being unveiled: at the time of writing, thalidomide sales are reaching over 200 million dollars per year, mostly deriving from yet another off-label use for multiple myeloma, among other indications (Ashburn and Thor, 2004). The thalidomide story teaches us that a drug can be harmful in one patient population (pregnant women) and highly beneficial in another. Correctly identifying the molecular processes affected can help to predict adverse effects and reorient the drug accordingly. The root cause of the discovery is here again serendipity, yet a better understanding of the proteins interacting with the molecule could have been helpful to predict the opportunity (Sampaio et al., 1991).

1.2.3 Raloxifene: expanding the application line

The last case discussed is less striking and unexpected than the two previous ones. Raloxifene (structure on Figure 1.3) is a selective oestrogen receptor modulator (often abbreviated as SERM) marketed as Evista by Eli Lilly. Similarly to other oestrogen modulators, such as tamoxifen, the original indication during preclinical developments was breast cancer (Ashburn and Thor, 2004). Despite early studies showing the positive effect of anti-oestrogens on osteoporosis in rats

in 1987 (Jordan et al., 1987), raloxifene's potential for this usage was not experimentally confirmed until 1994 (Black et al., 1994). Eventually, the molecule successfully passed clinical trials in 1999, with osteoporosis as a unique indication. However, the polypharmacology of the drug, particularly its action against breast cancer, was still under investigation. Finally, in 2007, the FDA approved raloxifene as a preventive agent for breast cancer in post-menopausal women (FDA, 2007), therefore extending the line of application of the drug back to its originally thought indication. In summary, raloxifene started its life as a breast cancer agent, was repositioned against osteoporosis, likely for strategic and commercial reasons, and eventually got approved for its breast cancer preventive properties. This drug is an example of smart and continuous development, expanding from one indication to another. The fundamental reasons behind the repositioning are grounded on early-stage experimental evidence and not due to a surprising effect appearing in clinical trials. The drug's polypharmacology was known and the indication of the molecule derived accordingly. Raloxifene is a good example of "educated repositioning" or indication discovery. The available information can help to appreciate and understand what the chemical might do from empirical evidence.

1.3 Computational approaches towards drug repositioning

The theory and the clinical cases presented the reality of drug repositioning. I briefly commented on the fundamental reasons enabling new usages and stressed the importance of serendipity in this process. Now the fantasy of many scientists working in the drug discovery domain is to be able to formally predict such repositioning scenarios and unveiling new pharmacology in an automated fashion.

In order to reach this distant goal or at least get closer to it, several computational approaches have been developed throughout the years. This section summarises the previous work done on the topic and motivates the novel way I explored, based on a formal representation of the *mode of action*. I chose to classify the different computational approaches based on the biomedical concepts used as the centre of the methodology. Some recent reviews (Ashburn and Thor, 2004) (Dudley et al., 2011a) (Hurle et al., 2013) of the field can also provide com-

plementary information to the interested reader, as well as a different perspective and logical structure of the topic.

Abstracted to the extreme, the goal of a repositioning initiative is to establish a link between a drug and a disease. This edge represents the indication or a prescription possibility for the molecule. In order to computationally forward new indication hypotheses, it is possible to use the biomedical concepts depicted on Figure 1.4 as proxy.

These various biomedical concepts represent different abstraction levels over a biological system, more or less close the biochemical reality and on the top of which it is possible to derive a value for computation or comparison. Usually, a similarity metric is derived from the property studied (e.g. chemical structure or gene expression level), which serves as a descriptor to rank the information and predict the new indication, materialised by a new link between a drug and a disease or a molecular target. Approaches can be roughly divided into groups, named after the central property of the analysis. Some alternative methods rely on a combination of concepts and are presented at the end of this section.

The abstraction process and its relation to biomedical knowledge will be further discuss in Chapter 2 and with the black box machine model (section 2.2.1.2).

1.3.1 Chemical structure-based approaches

Traditionally speaking, orally active drugs are mostly small lipophilic molecules (Lipinski et al., 1997). It therefore intuitively makes sense to directly look at the chemical structure to compare the similarity among drugs: similar structures are deemed to lead to similar biological outcomes. This rule of thumb goes by the name of *similar property principle* (Johnson and Maggiora, 1990) and is at the core of any quantitative structure-activity relationship (QSAR) study. A variety of methodologies exist to calculate the structural similarity between two chemicals, such as fingerprints or clustering algorithms (Eckert and Bajorath, 2007). These methods can be used to perform ligand-based virtual screenings: from a set of known active ligands, trying to find in a database of interest the structurally related molecules, supposedly bioactive too.

In the context of drug repositioning, one can search only among approved compounds for instance. This approach was successfully used by Noeske et al. (2006) implementing an unsupervised machine learning algorithm (self-organising

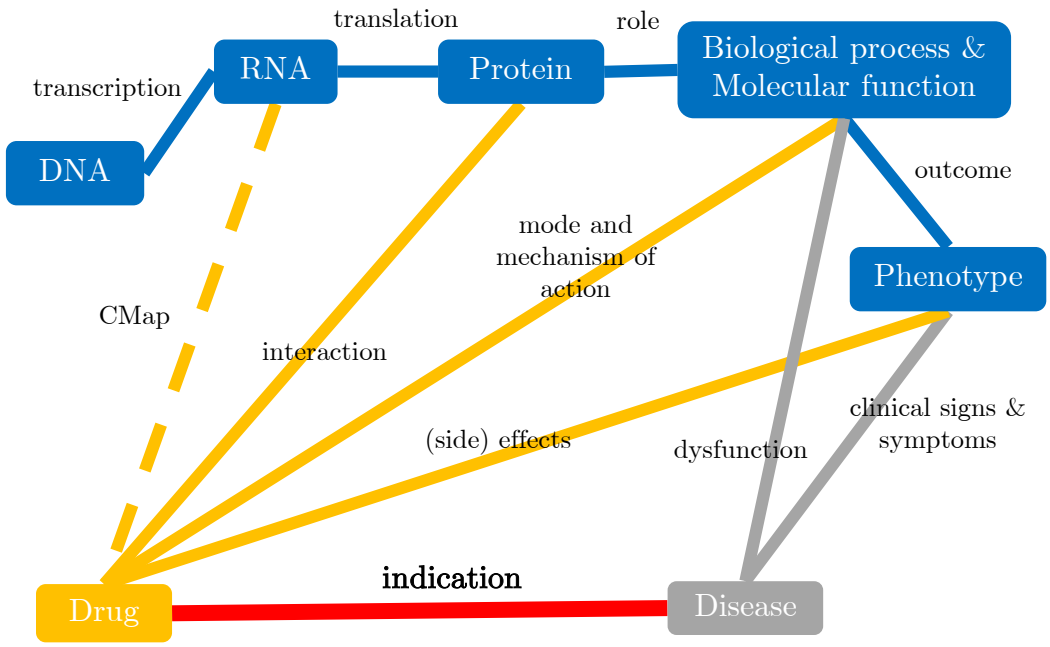


Figure 1.4: Conceptual map of the relationship between the different biomedical concepts. Relation related to the drug and its action are in orange, diseases in grey and biological concepts are in blue. Computational drug repositioning methods are based on either one or a series of such concepts in order to forward new indications for a drug, ultimate goal (red edge).

map) in order to cluster chemicals based on their structure. Molecular scaffolds were represented as vectors and used as input for the clustering step. The authors identified cross-activities for the metabotropic glutamate receptor antagonists, on other protein targets such as the dopamine D2, histamine H1 and muscarinic acetylcholine receptors. The off-target predictions were experimentally validated *in vitro* and shown to be active, yet not necessarily pharmacologically relevant due to weak binding. The new knowledge on off-target binding from this study can lead the way to potential new usage for the drugs, by further modifying and optimising the molecular structure for instance.

Another interesting approach related to structural similarity for off-target identification comes from the work done by Keiser et al. (2009). For this project, known ligands were grouped based on their known target binding partners and chemical features. The method is called *similarity ensemble approach* and *calculates whether a molecule will bind to a target based on the chemical features it shares with those of known ligands, using a statistical model to control for*

random similarity (Lounkine et al. (2012) - adaptation of BLAST for chemical structures). In the case of drug repositioning, the molecules tested were only approved drugs. The results revealed a series of off-target cases from the similarity analysis. A retrospective investigation showed the validity of the approach; then some predicted off-target bindings were experimentally validated, providing insightful clues about the pharmacological mechanism of some drugs. In some cases, such as for fabahistin, the off-target affinity (5-HT5A) was even better than for the known canonical receptor (H1), opening doors for meaningful alternative indications (see Figure 1.5).

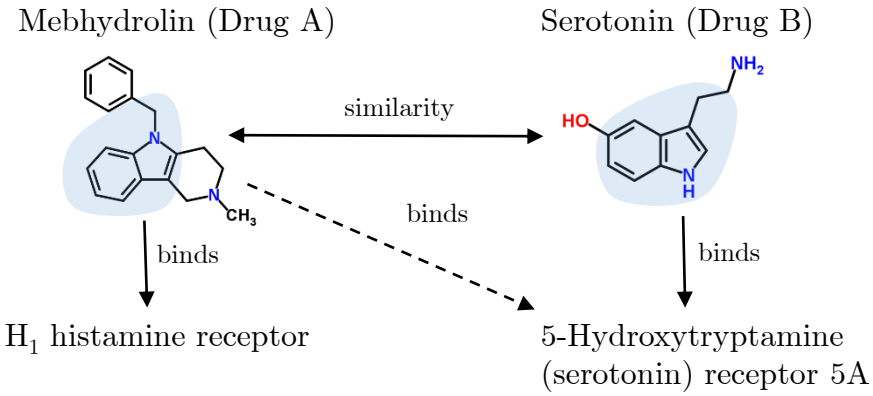


Figure 1.5: Drug repositioning using the chemical structure. Compounds with similar structures have similar biological activities (similarity principle). Drug A shares some similarity with molecule B, indicated by the blue areas. This observation leads to the conclusion that molecule A could be active on the canonical target of molecule B, and indicated accordingly. See Keiser et al. (2009) for full details on example. Structures obtained from the ChemSpider database.

Chemical-based approaches are intuitive and build on the accepted *similar property principle* (see Figure 1.5 for summary). However in practice, small changes made to a molecular structure can lead to drastically different biological outcomes. Moreover, the predictions made by the various methodologies have little overlap between them, stressing the difficulty to select the adequate one for the right scenario (Eckert and Bajorath, 2007). Some compounds also undergo chemical modifications by the cell before being pharmacologically active, therefore the structure as recorded in databases can compromise the value of a predictive statistical model. Finally, by using potent and optimised molecules as

starting point to infer new ligands or to train a model, there is a high risk that the predictions will manifest weak experimental pharmacology for the new indication. Indeed, such inferred compounds are not "dissimilar" enough to have an unexpected binding conformation with the protein and any differences in structure against the endogenous ligand will only reduce potency.

1.3.2 Gene expression and functional genomics-based approaches

Living systems can be understood by looking at the behaviour of their gene expression in a particular setting. Depending on the state of the system, certain genes are going to be over or under expressed, identifiable from the relative number of their messenger RNA (mRNA) molecules transcribed. Differentially expressed genes can serve as a proxy to characterise a molecular effect, so called the gene expression signature. This type of experiment is usually performed on a microarray, containing probes for the genes of interest (see Figure 1.6A). The approach provides a straightforward read of the condition studied and has been successfully used to find new indications for marketed drugs. In particular, the Connectivity Map (Lamb et al., 2006), was behind most recent repositioning stories (see Figure 1.6 for summary of the method).

The idea behind the CMap states that the action of a drug can be captured and compared by looking at the gene expression profile resulting from its administration onto a biological system. In this regard, the initiative recorded the molecular signatures of 164 FDA-approved molecules over 5 different cancer cell lines. The data is freely available and can help to perform various type analyses, for instance related to the understanding of the molecular mechanism of a drug.

Closer to my concerns, Iorio et al. (2010) used the resource for drug repositioning, by comparing small molecules on the basis of their CMap gene expression signatures: compounds with similar signatures were assumed to be functionally related, as they perturb the cell in a similar fashion. On this basis, the researchers identified communities of drugs sharing known protein targets and mechanism of actions. Interestingly, inside these hubs, most of the drugs were also sharing the same or similar therapeutic indications; yet outliers were present. Such cases were interpreted as repositioning opportunities, as these compounds appeared similar on the signature level (same gene expression profile), yet were clinically used for

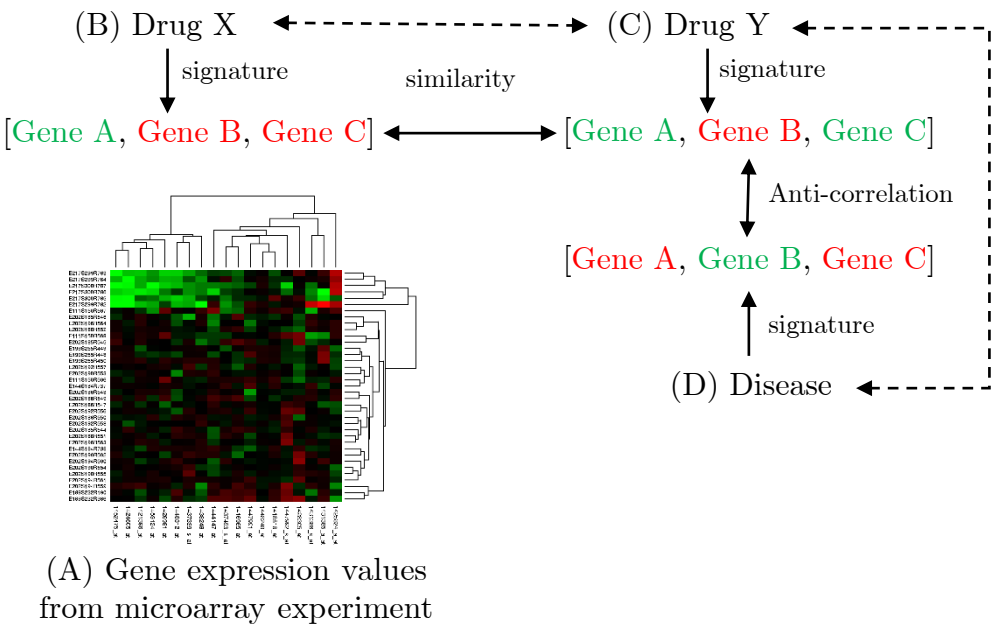


Figure 1.6: Drug repositioning using gene expression. (A) Example of result obtained from a gene expression experiment. Some of the probed genes are up-regulated (green), some of them down-regulated (red). (B) and (C) The gene expression data from the Connectivity Map provides a signature which can relate drugs on their functional aspect. For instance drug X and Y are considered similar because they share a significant amount of genes up and down related. (D) An analogue reasoning can be made with the relation drug-disease: disease signature can be treated by drugs with an anti-correlating signature.

different purposes. Based on these considerations, the authors discovered that fasudil, a potent vasodilator, can also be used as an enhancer of cellular autophagy. The prediction was experimentally verified on standard cell assays. The novelty of the methodology lies in the way variable gene expression data from different cell lineages was integrated in order to derive meaningful metrics.

Messenger RNA expression can reflect the activity of a drug, but it can also be used to characterise disease states. Following this assumption, Sirota et al. (2011) used a set of experiments from the Gene Expression Omnibus (GEO) in order to capture disease signatures from gene expression profiles. They further integrated this data with similarity values between drugs, derived from the CMap. The authors were able to find clusters of related diseases, appearing to negatively correlate with the signatures of the drugs currently used to treat them. The anti-correlation was used to predict the anti-ulcer drug cimetidine as a candidate for

the treatment of lung cancer. The efficacy of the molecule for the new indication was demonstrated *in vitro* and *in vivo* on a mouse model. The analysis also identified topiramate, currently used as an anticonvulsant, as an active agent for the treatment of inflammatory bowel disease (Dudley et al., 2011b), for which no cure currently exists. The relevance of the pharmacology in respect to the new usage was extensively confirmed *in vivo* in a rodent model. This study highlights a powerful feature of transcriptomics: even when little molecular information is known about the exact underlying pathology, gene expression signatures can help to efficiently abstract away from mechanistic details and correctly identify potential treatments.

A homologous approach has been used to overcome cancer cell resistance in leukaemia (Wei et al., 2006). Glucocorticoids are usually administered as treatment, yet sometimes the cancer cells of some patients develop a resistance to them. The gene expression profile of cells sensitive to the traditional treatment was compared against the signature of resistant cells, in order to characterise the molecular differences. Then using CMap data, the authors predicted rapamycin as a novel agent to overcome the resistance, as the two gene expression profiles were correlated. The prediction was experimentally confirmed, providing further insight regarding the new mechanism of action. Rapamycin, an immunosuppressant agent indicated to prevent rejection in organ transplantations, might find here a new life for the treatment of lymphoid malignancies. A similar methodology, based again on anti-correlation of disease state against drug signature from the CMap, connected ursolic acid to skeletal muscle atrophy (Kunkel et al., 2011). The molecule improved the muscular mass and reduced adiposity in a clinical study on humans and mouse.

Gene expression analysis has led to numerous success stories, thanks to the valuable resource that is the CMap. The technique does not require much prior knowledge about the action of the drug or the pathology behind a phenotype; the creation of signatures from direct mRNA readout helps to simply retrieve unexpected drug-disease associations. Another important lesson from transcriptomics and the likely reason behind these successful results is the functional aspect: drugs are solely characterised based on their role and action in the biological system, represented by a gene expression signature. The chemical structure is disregarded and almost irrelevant for such analyses.

Despite impressive results, the technique suffers from drawbacks, currently

subject to improvements (LINCS, 2014). First, the expression profile of the drug or disease must be available. The CMap provides a relatively small list of molecules. This collection is far from being representative of all approved and experimental drugs, limiting the compounds that can be investigated. Second, gene expression profiles can arguably define a disease state or a drug response; tissues are not considered in the CMap, the resource was only developed by recording the response of cancer cells, and is not necessarily relevant for all disease types. Finally, transcriptomics data also present considerable challenges in terms of statistical analysis, as recognised by the authors of the CMap project (Lamb et al., 2006).

1.3.3 Protein structure and molecular docking-based approaches

Most bioactive small molecules mediate their effects by interacting with proteins. This interaction can be analysed using computer software modelling the three dimensional (3D) structure of the target and the drug. This practice is known as molecular docking and commonly applied within drug discovery pipelines; the method helps to identify and optimise binding affinities in the active site of the target (e.g. pocket of an enzyme) in order to increase the potency of the drug developed (Haupt and Schroeder, 2011). Because of the popularity of molecular docking, it is not surprising to find drug repositioning attempts based upon this approach. As most of the compounds are known to interact with more than one protein (see section 1.1.1), the aim is to identify these potential off-targets, by screening against the 3D structure of proteins present in a given database. If the predicted off-targets are disease relevant, then the drug could be repositioned accordingly.

In this respect, a series of recent studies focused on binding sites and compared their relative similarities (Haupt and Schroeder (2011) - see Figure 1.7 for a summary). Looking only at the structure of protein's active sites guarantees to be as close as possible to the biochemistry and physical reality of the interaction. From over 6,000 binding site structures De Franchi et al. (2010) identified the synapsin I, protein involved in the regulation of neurotransmitter release as a new target of the drug staurosporine, known to bind the Pim-1 kinase (De Franchi et al., 2010). The finding was experimentally verified *in vitro*, yet the pharmacological

relevance of the new target remains to be shown.

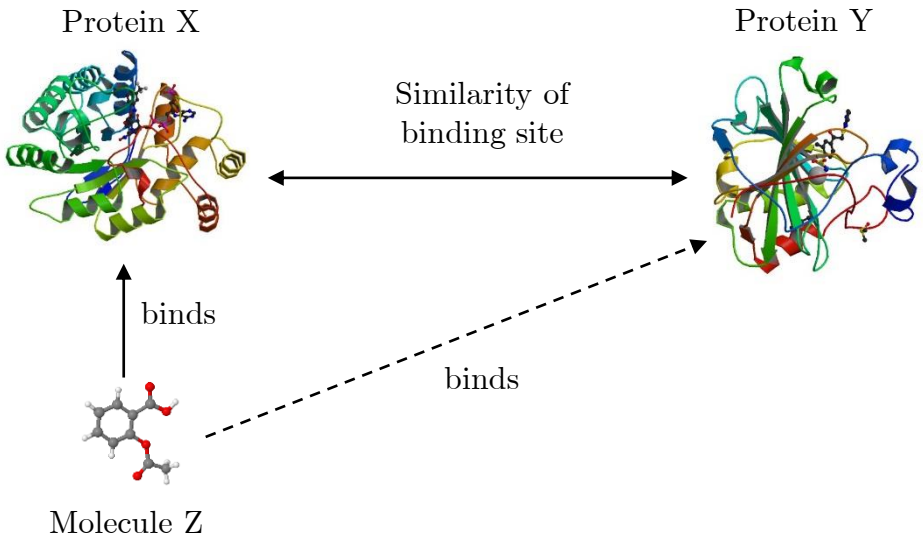


Figure 1.7: Drug repositioning using protein structure and binding site. The 3D structure of proteins and their respective binding sites can be compared using a scoring function. On this basis, it is assumed that similar binding sites can bind the same ligand. For instance, knowing that protein X has a similar binding site to the one in protein Y, and that molecule Z binds to protein X, one can forward the hypothesis stating that molecule Z should bind to protein Y too. Illustrations from the Protein Data Bank.

Zahler et al. (2007) performed an inverse screening (docking one compound over multiple binding sites) to characterise the off-target binding landscape of kinase inhibitors. This class of drugs, largely used in cancer therapy, has a notorious "promiscuous" behaviour. The virtual screening revealed a new enzyme, PDK1, as an off-target of indirubin. Here again, the prediction was validated *in vitro* via a phenotypic cell proliferation assay demonstrating the validity of the approach and providing insights regarding kinase inhibitors' side-effects.

Finally Kinnings et al. (2009) addressed drug resistant tuberculosis using molecular docking methods. The bacteria behind the condition can indeed sometimes resist the first-line drugs, and in such cases, the eradication of the pathogen becomes difficult. The authors followed a workflow called *selective optimisation of side activities* (SOSA), a technique developed to progressively move away from the original indication and optimise a compound across protein families (Wermuth, 2006). Briefly, the methodology is composed of the following steps: binding site extraction from 3D structure of protein sequences, identification of

similar binding sites across the proteome using a search algorithm, and finally manual docking analysis to make sure the physical interaction is possible. From this pipeline, the group predicted the approved drugs entacapone and tolcapone (prescribed for Parkinson's disease) to be potent against the enoyl-acyl carrier protein reductase, an enzyme essential for the synthesis of fatty acid in *Mycobacterium tuberculosis*. The drugs were experimentally proven to be active *in vitro*, using commercially available tablets. Even more interestingly, the new mode of action introduced with these two compounds could bypass the drug resistance encountered in *Mycobacterium tuberculosis*, and provide a valuable treatment for affected patients.

Despite the mentioned successes, molecular docking strategies for drug repositioning suffer from drawbacks. First, 3D structural data must be available. Databases such as the Protein Data Bank (PDB) contain numerous records, however they are still very far from covering the whole proteome (Haupt and Schroeder, 2011). Secondly, it can be challenging to automatically recognise a binding site, in particular when the protein structure was crystallized without the ligand. Finally, as all methodologies generate a large number of false positives, experimental and manual validations are the only solution to evaluate the predictions. Sometimes a single amino acid difference will totally change the pharmacology of the binding site (Kruger et al., 2012), a difficult problem to handle when the structures are analysed and aligned in an automated fashion.

In conclusion, protein-based approaches are arguably the closest methodology to the actual physical interaction between a drug and a protein target. Docking approaches provide a detailed low level picture of the biochemical complex, yet still suffer from challenges on the modelling side. The identification of off-target proteins does not necessarily always yield repositioning opportunities, and the results always have to be interpreted in a broader biological context.

1.3.4 Phenotype and side-effect-based approaches

The phenotype can be defined as *the set of characteristics or traits attributed to an organism*. Examples of phenotypes are the morphology, developmental, biochemical or physiological properties (Wikipedia, 2014f). This concept is widely used in biological sciences, to express the high level observations made when looking at a living organism. The phenotype is arguably the most primitive interaction

between the biomedical scientist and its object of study: while travelling across the world, Darwin built the evidence for evolution from the phenotype of barnacles (Darwin and Bynum, 2009). Gregor Mendel first described inheritance based on the traits observed in pea plants (Mendel, 1866). None of these scientists had any idea about the actual molecular mechanism responsible for the observed patterns, yet their phenotypic observations were strong enough to forward valid conclusions. This exercise is still very commonly applied in clinical settings. Every time a doctor diagnoses a patient, he or she primarily relies on a phenotypic characterisation of the signs and symptoms present in the patient. As mentioned earlier in this chapter (section 1.1.1), phenotypic-driven screenings are also routinely performed within drug discovery pipelines. It actually appears to be the best technique to bring new medicine to market, according to a recent study (Swinney and Anthony, 2011). This successfulness can be attributed to the fact that a phenotypic observation is a more accurate representation of the underlying system; the physiological context is preserved, as opposed to target-based assays, therefore *in vitro* lead compounds have better chances to stay active when scaling to animal models and eventually clinical trials (Duran-Frigola et al., 2012).

Back to drug repositioning concerns, side-effects can also be seen as phenotypes. The sildenafil story emphasises their importance: no matter how potent a drug is in animal model or in *in vitro* assays, its true pharmacology will only appear during the clinical trials. Accurately characterising these side-effects can help to reposition a drug or reveal new interaction partners, as highlighted by two studies, summarised here (see Figure 1.8).

Drugs with similar target binding profiles cause similar side-effects (Fliri et al., 2005) (Fliri et al., 2007). Starting from this rationale, Campillos et al. (2008) defined the side-effect profiles for approved drugs, then used similarity among these to identify the drug's off-targets. The side-effects were first extracted using text-mining from package inserts, in order to build a statistical model informing about the likelihood for two drugs to have a common target. The authors then focused on compounds from different therapeutic categories, yet with a high probability of sharing a target according to the model. They experimentally tested 20 of such predictions, validating 13 of them, 11 with an inhibition constant under 10 micromolars. The originality of the method demonstrates the molecular relevance of side-effects, and their potential to identify off-targets and reassign a therapeutic molecule to a new indication. An interesting aspect of this approach is the rep-

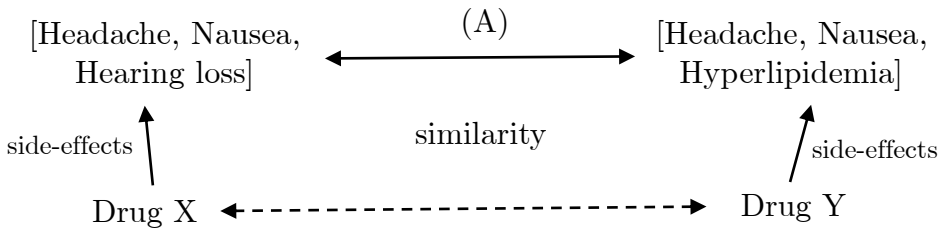


Figure 1.8: Drug repositioning using phenotype information. Knowledge about the phenotypic outcome triggered by a drug can be used in order to establish relative similarities. (A) The diagram illustrates a theoretical example using reported side-effects: the more side-effects are commonly shared by two drugs, the more similar these two drugs are. The similarity can be used to either derive potential off-targets or new indications.

resentation of side-effects. Just as any phenotypic manifestation, words or terms, derived from observations, are still the best way to express them. For this study, the authors used the Unified Medical Language System (UMLS) (Bodenreider, 2004), a controlled vocabulary provided by the National Health Institute. From the experimental validation shown by the research group, one can infer that ontologies and controlled vocabularies indeed have the potential to generate sound predictions.

Another approach was presented by Yang and Agarwal (2011). The authors used side-effects from the database SIDER (Kuhn et al., 2010) to link diseases and extract drug repositioning opportunities. Chemicals were linked to pathologies using the information available in pharmacogenomics knowledge base (PharmGKB) (Whirl-Carrillo et al., 2012). The approach values evidence showing that drugs used to treat similar diseases have similar side-effects. In this respect, side-effects can be indicators of a common underlying mode of action, and two drugs sharing a significant number of side-effects can be used to treat the same pathology. Sometimes a side-effect can also be deleterious in one case and beneficial in another. For instance hypotension is usually thought of as an unfavourable side-effect, yet the compounds producing such an action could be used as antihypertensive agents. Following this hypothesis, a predictive model (Naive Bayes) was then built in order to assign drugs to diseases on the basis of side-effects. The high

performance of the evaluation demonstrated the relevance of the methodology. The authors further developed a tool to predict new indications for compounds under development in a systematic fashion. No experimental validation was presented, yet a robust evaluation of the methodology was performed.

Fifty years ago, the phenotype was at the centre of drug discovery. With the advent of molecular biology, it was progressively replaced by the target-based paradigm (Duran-Frigola et al., 2012). Phenotype-based approaches are still interesting for drug discovery, as they report the effect of a given substance on the level of whole organisms, which may be more representative for clinical applications. Thanks to the recent development of ontologies and methods (Hoehndorf et al., 2007) (Hoehndorf et al., 2011b), it seems that the biomedical community has now better way to record, capture and align phenotypic information.

1.3.5 Genetic variation-based approaches

Closer to the molecular level, genetic variations can also provide valuable insights regarding drug repositioning opportunities. Due to the recent implementation of high-throughput DNA sequencing methods and analysis pipelines, it indeed becomes increasingly cheaper to sequence individuals and study their genotypes. From the information generated, one can isolate common mutations in the DNA that are significantly associated with a phenotypic trait. This method is known as genome-wide association study (GWAS) and is typically used to relate a single-nucleotide polymorphism (SNP) to a disease. The data about SNPs and their association to pathologies is indexed in databases, such as the one provided by National Human Genome Research Institute (`http://www.genome.gov/gwastudies/`). Using this resource, Sanseau et al. (2012) performed an analysis to unveil potential new indications for protein targets from GWAS. The logic behind the approach is that the association between a SNP and a trait from a GWAS can be extrapolated as a relation between a gene and a disease (if the only traits considered are diseases - see Figure 1.9). Then knowing that a drug targets the given gene product, one would expect for the indication of the drug to be the same as the trait studied in the GWAS. For instance, in the gene encoding 3-hydroxy-3-methylglutaryl-CoA (HMGCR), a SNP was significantly associated with the trait *LDL cholesterol* (Kathiresan et al., 2008). A class of drugs, the statins, are known to target this gene product and

are indicated as cholesterol lowering agents (hypercholesterolemia). The authors identified 97 of such cases, where the SNPs support the current drug indication and give more confidence to the biological role of the protein. On the contrary, for 123 associations, the authors reported a mismatch between the trait associated with the gene and the current indication of the drug; these associations have been inferred as repositioning opportunities. For example, denosumab is a monoclonal antibody indicated for the treatment of osteoporosis and bone cancer. Its main target, the protein TNFSF11 (tumour necrosis factor superfamily, member 11) contains a SNP paired with Crohn's disease (Franke et al., 2010). Based on this evidence, denosumab could be tested for this later condition. Another example reported by the authors is nepicastat, a small molecule indicated for cocaine addiction and post-traumatic stress disorder. The target of the compound, DBH (dopamine beta-hydroxylase), has been associated with the trait smoking cessation in a GWAS (Furberg et al., 2010). This result suggests a new and unreported use for nepicastat, as a drug for smokers willing to stop.

The methodology presents some pitfalls, as shown by the prediction made for NOS2 (nitric oxyde synthase 2) inhibitors to be active against psoriasis; clinical trials unfortunately failed to show any significant effects. The gene-disease relation is indeed complex in practice, and sometimes more information is needed to appreciate the potential effect of a drug. Moreover, since GWAS does not provide any information regarding the direction of the pharmacological effect and it is difficult to know for instance whether an agonist or antagonist should be used to produce an outcome. Despite these restrictions and because of the impressive recent progress made in genome sequencing, this approach, or a related methodology, might gain in importance in the coming years.

1.3.6 Disease network-based approaches

Traditionally, diseases have been grouped together, on the basis of the cause of the pathology (e.g. infection) or the biological dysfunction observed (e.g. uncontrolled cell growth) for instance. As similar diseases are treated in a similar fashion, a better characterisation of the relation holding between pathologies can generate drug repositioning hypotheses. I will briefly present some of the work done in this direction, on the construction of a "diseasome" or network of relationships between diseases (see Figure 1.10 for summary).

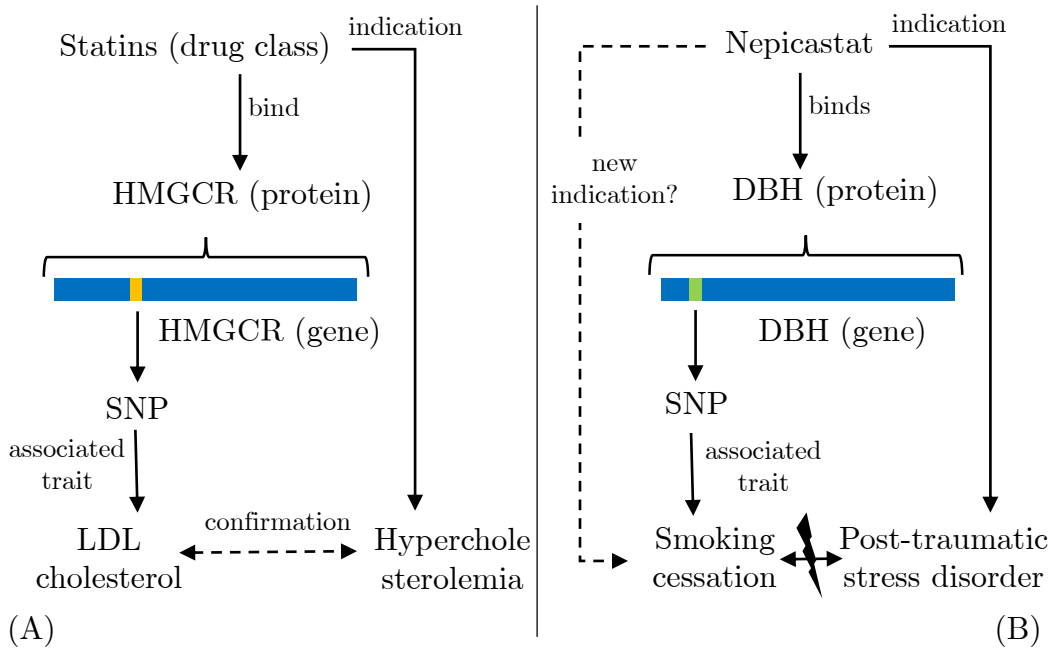


Figure 1.9: Drug repositioning using genetic information. (A) Single-nucleotide polymorphism (SNP) are associated with a phenotypic trait, here LDL cholesterol. The gene where the SNP is found (HMGCR) encodes for a protein, targeted by statins (drug class). Statins are indicated as cholesterol lowering agents, which is confirmed by the trait associated with the SNP. (B) Sometimes the trait associated with the SNP diverges from the indication of the drug, as shown on the diagram (post-traumatic stress disorder against smoking cessation). In such cases, a repositioning hypothesis can be generated. Examples are detailed in the text. See Sanseau et al. (2012) for more explanations.

Chiang and Butte (2009) defined diseases from the list of drugs used in their therapies and off-label indications. Despite being fairly simplistic, the rationale is backed by successful examples and commonly performed in clinical settings. The authors performed an associative indication transfer, namely, given two similar diseases, proposing to use a drug indicated only for one of them as a therapy for the other. From 700 diseases and 2,000 drugs, over 150,000 new associations were generated. Interestingly, the new indications are in agreement with clinical trials data, the predicted new usage has often been reported by doctors (12 fold enrichment against random). For instance, atorvastatin, a cholesterol lowering agent, was predicted to be active for asthma, Crohn's disease and myocardial infarction; all these associations have been positively reported in clinical trials, proving confidence in the methodology. For the same drug, some of the new

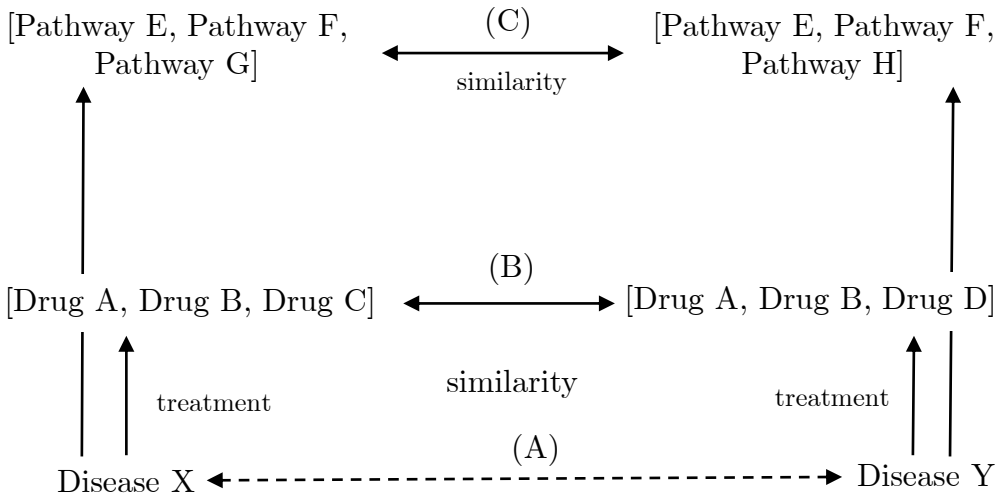


Figure 1.10: Drug repositioning using disease relationships (diseasome). The similarity between two diseases (A) can be calculated by looking at the shared drugs used for the treatment of these diseases (B) or at the commonly shared pathways (C). When applied to all diseases, one can build a "diseasome" or disease map, useful to relate indications and find drug repositioning opportunities.

associations have no clinical evidence, such as activity in breast cancer and osteosarcoma. Accordingly, it is possible to investigate the action of the drug for these pathologies. This work illustrates one possible approach relating diseases; two other methodologies presented now respectively construct the network from shared pathways and functional modules. The first one (Li and Agarwal, 2009) built a map linking diseases from public resources (e.g. Reactome, Kegg pathways and text-mining). Diseases with commonly deregulated pathways were deemed to be similar. The properties of the resulting graph were analysed and the authors showed how their work can provide new insights about disease relationships. No analysis for repositioning opportunities were performed, yet the map can serve as a starting point to identify similar conditions, on which one can perform an indication transfer as before. Finally, Suthram et al. (2010) constructed a disease graph from gene expression profiles and protein networks. An analysis revealed 59 functional modules, shared by half of the diseases studied. These modules relate pathologies on their molecular basis and help to better understand the internal wiring of the system. Similarly to other methods, once the disease network is created, drug repositioning hypotheses can be generated. As a conclusion, despite

not directly addressing drug repositioning, disease maps can provide valuable insight regarding the usage of a drug. Such approaches also question the current way of classifying diseases, by considering molecular information as signature or definition.

1.3.7 Machine learning and concepts combination approaches

The approaches presented before mostly focus on one of the concepts of the map shown on Figure 1.4 and orient their analysis around it. It is also perfectly possible to use a combination of these biomedical descriptors to train a machine-learning algorithm and then generate predictions out of the statistical model (see Figure 1.11). Two recent studies address drug repositioning from this perspective. In both cases, first a series of biomedical heuristics is defined, then the model is trained on known data and predictions are made.

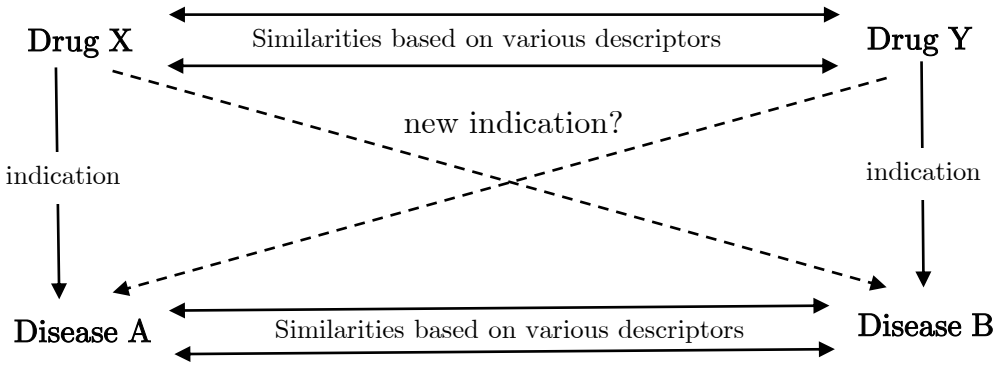


Figure 1.11: Drug repositioning using a combination of descriptors. A machine learning algorithm is trained over a series of features, such as chemical similarity, shared target proteins, etc. After evaluation of the model, some repositioning predictions can be generated from the statistical learning.

The first method presented is called PREDICT (Gottlieb et al., 2011). In order to train the machine-learning algorithm, the authors decided to represent separately drug-drug and disease-disease associations. Drug-drug associations were characterised from fingerprinting of their chemical structure and the reported and predicted list of side-effects. These drug-drug associations were further enriched

with information related to the targets of the drugs: the sequence similarity, distance in the protein-protein interaction network and semantic similarity of their GO annotations. The disease-disease associations were more simply characterised based on their semantic similarities calculated over the Human Phenotype Ontology (HPO) from the annotations present in the Online Mendelian Inheritance in Man (OMIM) database. From these gold-standard associations, the authors trained a logistic regression classifier to recognise real associations from fake ones. The model was evaluated against the predictions made by other methodologies, such as guilt-by-association and CMap approaches (Lamb et al., 2006), presented earlier in this chapter. The evaluation shows little overlap between the various methodologies; it is difficult to align the various datasets, as often the diseases and drugs considered are different. Some drug repositioning predictions were then generated and evaluated from clinical trials data. Around a third of the predictions were reported as already investigated, giving confidence in the outcome of the methodology. In the last part of their work, the authors substituted the disease-disease associations based on phenotypic similarity with gene expression profiles. The idea behind this step was to test the developed method for personalised medicine: assuming one has access to the gene expression profile of a patient, can PREDICT find the best drug to administer to the individual? Results were encouraging; the method reports high recall and specificity (area under curve of 0.92 obtained from receiver-emitter curve), providing a tangible proof-of-concept for the algorithm.

The second method presented (Napolitano et al., 2013) is very similar to PREDICT. The main difference comes from the machine-learning methodology, which was Support Vector Machine (SVM) in this study. The algorithm was used to predict therapeutic categories of the Anatomical Therapeutic Chemical Classification System (ATC), misclassification being re-interpreted as drug repositioning hypotheses. The researchers also incorporated structural similarity, protein-protein interaction network distance and gene expression data as starting features to train the SVM. After standard machine-learning evaluation procedures, the authors derived repositioning predictions. The main hypotheses were anthelmintics compounds to be active as antineoplastic agents and antineoplastic drugs to be used as systemic antibacterials.

Machine-learning based approaches to drug repositioning provide a way to combine various descriptors into one statistical model, with the aim of increasing

the accuracy of the predictions. The techniques presented however face some important pitfalls. One of them is the interpretation of the repurposing hypotheses: the statistical model is a black box, hiding the rational evidences of why a compound was chosen. Most of the hypotheses end up being evident cases, easily explainable for a biologist by looking at the chemical structure or known off-targets of the compounds. This result is maybe due to an over-training of the machine. Finally, one can question the biomedical meaningfulness of integrating a large number of descriptors; since diseases are subtle and unique, too much information risks blurring the outcome by overlooking important biological mechanistic details.

1.3.8 Summary

A variety of approaches have been tried in order to computationally repurpose drugs. The field is still in its infancy, as revealed by two factors.

First, it is still not clear which method provides the best results and why. The only absolute way to evaluate the predictions is when a drug will be routinely indicated in the clinic from a hypothesis generated in silico; as far as I know, there is no compelling story illustrating this yet. It is not surprising; developing a new drug is a long process, taking over twelve years to achieve and trapped with legal and economic hurdles. Knowing that the first study reported by PubMed for the keyword search "computational drug repositioning" dates to 2006 (An and Jones, 2006), or 7 years before the time of writing, it seems realistic not to find any clinical examples yet.

Secondly, each method addresses the drug repositioning problem from a different angle or biomedical concept, which complicates the evaluation process. Objectively integrating the results from the various approaches is a difficult task, as the starting datasets are about different molecules and diseases and produce different type of outcomes. It would be beneficial for the community to have a standard dataset, covering the legal indications, as well as the known and confirmed alternative ones. Computational methods could use such a resource to benchmark their performance and evaluate their predictive power and perform error analysis. Immaturity also means creativity in this case, illustrated by the numerous methods implemented. Table 1.1 and 1.2 provide a summary of the approaches presented, alongside their respective strengths and weaknesses.

Biomedical concept	Rationale	Biomedical advantages	Technical advantages	Biomedical pitfalls	Technical pitfalls	References
Chemical structure	Similar chemical structures have similar biological outcomes (similar property principle)	Off-target identification, straightforward interpretation, can be used on new chemical structures with unknown activities	Fast algorithm available to encode and search for chemical structures (fingerprint), large number of structures available	Prediction can have weak binding, not necessary pharmacologically active, small changes made to a molecular structure can change a lot the biological outcome, compounds may undergo chemical modifications by the cell (pro-drugs), therefore original structure is not so helpful	Chemical structures information in databases is sometimes erroneous	Noeske et al. (2006), Keiser et al. (2009)
Gene expression (mRNA)	A biological state can be defined by the list of genes under or over-expressed (signature) in the given state. It is possible to define disease states and drug states (CMap) and analyse their relative similarities.	Systematic characterisation of the biological function, no previous knowledge required about diseases or protein targets	Assay well understood and cheap, CMap data freely available and extending	Too simplistic to characterise some states, important mechanistic aspect can be overlooked	Selection of the representative genes is challenging, CMap data were recorded on cancer cell, might not feat all types of diseases or it can bias the signature	Iorio et al. (2010), Sirota et al. (2011), Dudley et al. (2011b), Wei et al. (2006), Kunkel et al. (2011), Lamb et al. (2006)
Protein	Computational modelling of the physical binding of a drug to a protein	Off-target identification, straightforward interpretation, can be used with chemical structures with unknown activities, close to biochemical reality	Numerous tools available to perform docking studies	Predicting a binding is challenging because of molecular dynamics, high number of false positive predictions	Structural databases are not complete, not all proteins are considered	Ashburner et al. (2000), Haupt and Schroeder (2011), De Franchi et al. (2010), Zahler et al. (2007), Kinnings et al. (2009)

Table 1.1: Summary of drug repositioning approaches.

Biomedical concept	Rationale	Biomedical advantages	Technical advantages	Biomedical pitfalls	Technical pitfalls	References
Genetic	Genomic identification of phenotypic traits	Potentially closer to individual patients (SNP)	Large amount of SNP information already available and growing everyday, decreasing cost of sequencing	Does not provide mechanistic details nor context	Genome data analysis is still challenging	Sanseau et al. (2012)
Disease	Similar diseases receive similar treatment	Systemic characterisation and classification	Various methodologies can build the diseases map	Does not directly address drug repositioning	Highly rely on curated knowledge	Chiang and Butte (2009), Li and Agarwal (2009), Suthram et al. (2010)
Combination	Training of a machine-learning algorithm from a series of descriptors	Incorporation of multiple dimensions, providing a more consistent representation of the biological phenomenon	Numerous machine-learning program and methodologies available	Biological interpretation difficult (black box approach), risk of over training the system, does not provide mechanistic details	Choice of the heuristics is challenging	Gottlieb et al. (2011), Napolitano et al. (2013)
Biological process and molecular function	Formal representation of the mode and mechanism of action	Systemic characterisation of the biological role, mechanistic description	Large amount of information available, data integration	Predictions can have weak binding, not necessary pharmacologically active	Highly rely on curated knowledge	Croset et al. (2013b)

Table 1.2: Summary of drug repositioning approaches (continued).

To conclude, computational drug repositioning appears as a topic of growing interest in the scientific community, as inferred from Figure 1.12. A series of methods have been developed in the last 5 years, summarised and discussed in this chapter. Drug repositioning is a subset of a larger problem type: indication discovery and network biology (Hopkins, 2008). In can be summarised as taking advantage of our increasing knowledge about systemic behaviour to computationally design smart drugs. Various abstraction level can be considered, as shown by the series of biomedical concepts used as starting points (see Figure 1.4).

The conclusions drawn from this state-of-the-art review oriented my work and motivated me to explore drug repositioning from the perspective of biological processes and molecular functions. More precisely, I aimed at logically model the regulatory events to derive a functional classification of drugs.

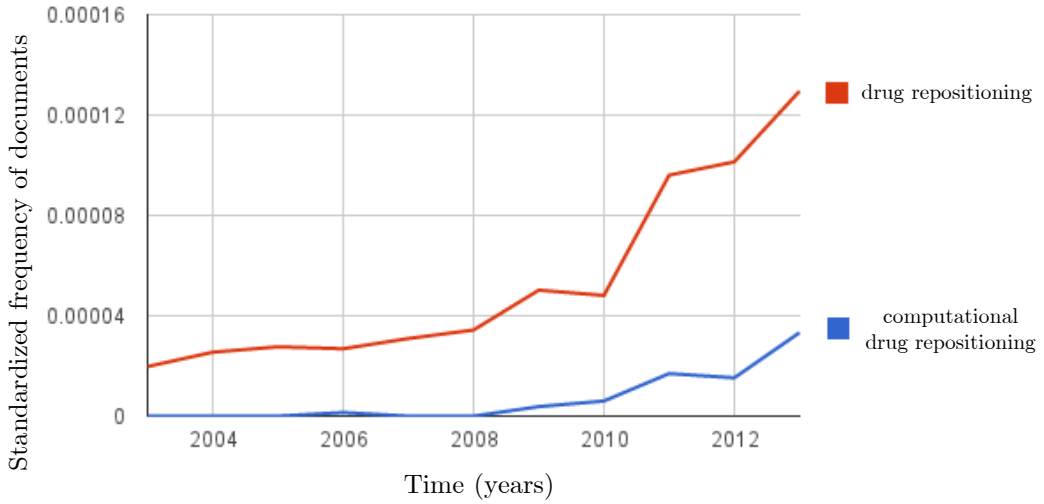


Figure 1.12: Evolution trend of the documents related to drug repositioning. Standardised frequency: number of documents indexed on PubMed for a search divided by the total number of articles published the same year. The higher, the more popular a topic is. The frequency increases with the time for both searches, showing a growing interest in the domain.

1.4 Thesis: biological process and molecular function for drug repositioning

The previous section (1.3) gave an overview of the various computational techniques developed for drug repositioning within the last decade. I believe that the

most promising results appear to come from gene expression analyses (section 1.3.2), driven by the large amount of public data available and encouraging *in vivo* results. Additionally, gene expression provides a functional insight about a biological state, valuable in practice: from a drug indication perspective, it is indeed much more pertinent to know what a molecule does (role or function) rather than what it physically looks like (structure).

Following this philosophy, I argue in favour of the *mode of action* (MoA), as a means to formally define the function of a drug. The concept brings unique assets, such as bridging from the molecular to the phenotype level, as well as providing a discrete mechanistic understanding of the role of a drug.

In this section, I will introduce my thesis, namely the vision and the rationale behind the approach of using curated knowledge of biological processes and molecular functions to characterise the drug repositioning landscape. Processes and functions provide a flexible abstraction layer in between biochemistry and systems biology. My thesis work is based on a computational characterisation of the mode of action of marketed drugs. This notion is particularly relevant to drug repositioning, as the concept provides mechanistic insights regarding the broad pharmacological action of a drug. Moreover, a compound can also have several modes of action, representing the various biological processes the molecule can perturb from its polypharmacology.

1.4.1 Rationale

Knowing the potential role a drug can play in a living organism such as a human body enables one to logically re-use the compound for a different indication. Moreover, a drug binds to multiple proteins, themselves involved into multiple biological processes (see section 1.1.1). Therefore a drug can potentially play a multitude of roles, or in other words can have several MoAs, which are accountable for its polypharmacology.

In practice, the biological roles of drugs are described as mechanism or mode of actions. The mechanism of action can be defined as the biochemical interaction that gives rise to the pharmacological effect of a drug. For instance, the term *phosphodiesterase-5 binding* is the mechanism through which the sildenafil molecule produces its action. Similarly, the MoA defines in a more abstract fashion the broad activity of the molecule on an organism. The terms *pro-penile erec-*

tion agent and *vasodilator* both are valid MoAs of the same sildenafil molecule. Another example of MoA is *anti-blood coagulant*, representing the capacity of a drug to inhibit or decrease the extent of blood coagulation (see Figure 1.13 for example).

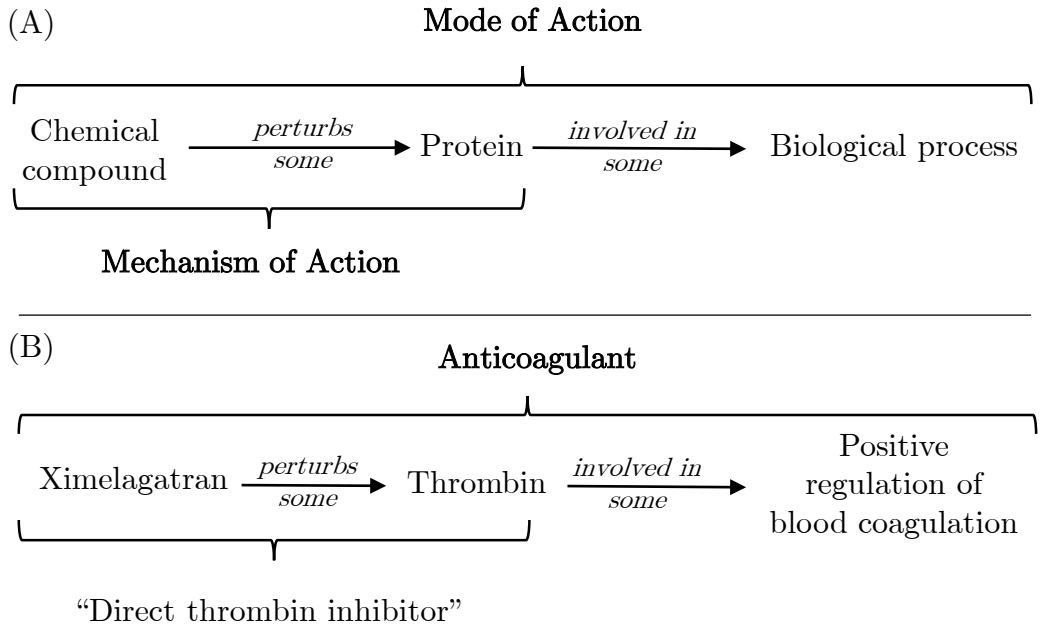


Figure 1.13: Schematic representation of the concept of mechanism and mode of action. (A) The mechanism of action can be defined as the physical activity of the ligand on a protein target. The mode of action characterises the pharmacology of the small molecule in the context of the organism. (B) Examples related to blood coagulation illustrating the usage of the mechanism and mode of action concepts.

The concepts of mode and mechanism of action are broadly used in drug discovery; they help to classify drugs into therapeutic groups. Even more importantly, a chemical is assigned as a treatment to a disease because it exhibits a particular MoA. One example is the case of high blood pressure, a common medical condition leading to an increased risk of heart attacks and strokes (NHS-Choices, 2014b). In order to chemically decrease the blood pressure, several biological solutions can be considered. A first approach could remove the excess of salt from the body, thereby decreasing the tension in the blood vessels. An alternative solution could inhibit the vasoconstrictive signalling of a hormone. Finally, it is also possible to act directly on the cells physically narrowing the vessels and preventing their unwanted action this way (Ong et al., 2007). Now if a chemical was

known to be capable of producing any of these biological actions, it would then be possible to indicate it for the treatment of the hypertension. This logic puts emphasis on the mechanistic details of the pharmacological effect. The MoA explicitly characterises the function of the drug in the cellular machinery, and helps to appreciate its overall effect. As opposed to statistical methods and black box techniques, the explicit reasons why a particular drug shows a particular outcome can be derived from curated knowledge in a systematic fashion.

A MoA-based approach comes with unique theoretical advantages. First, it provides a systematic high-resolution picture of the function of drugs, just like gene expression experiments. Moreover, this characterisation of the function is discrete and granular. Unlike gene expression data (CMap), the function of the molecule is not vaguely summarised in a gene expression signature. MoAs are discrete categories, and a compound can be put in many of them, as illustrated with hypertension. In this respect, the MoA can be seen as the equivalent of a protein's functional annotations (Ashburner et al., 2000), but for drugs. Finally, the MoA provides a unique level of abstraction over biological systems: it logically links the biochemical mechanistic interaction, as studied with docking and chemical similarity, to a high level phenotypic process, such as a disease or a side-effect.

1.4.2 Towards the specification, implementation and analysis

The computational use of the MoA to repurpose drugs comes with obstacles. I gave in the previous section some informal examples of MoAs: *vasodilator*, *anti-blood coagulant*, *anti-ageing agent*, etc. However, in order to provide helpful information, all MoAs (or a large number at least) have first to be unambiguously and formally defined. From these examples and Figure 1.13, the reader can see that MoAs are terms, boxes or categories; classical mathematical frameworks used in biology and bioinformatics, such as statistics, do not provide any means to address this concern.

A MoA describes logical events happening inside the complex cellular machinery: intuitively, an anti-blood coagulation agent is a type of compound capable of modifying the activity of a protein target somehow involved in the blood coagulation. Therefore representing the axioms or logical links underlying these

statements can set the basis to derive new MoA categories, as shown in Figure 1.13. However, the definition of MoAs from this perspective will necessarily rely on existing knowledge or data. From the example before, it is indeed necessary yet sufficient to know first that a drug binds a particular protein and secondly that this protein is involved in the coagulation process in order to formally derive the MoA of the compound.

The implementation of the rationale is described in the rest of this document and organised as follow.

1.4.2.1 Chapter 2 - Description logics and biomedical knowledge (Specification)

Before representing MoAs and using them to classify drugs, the specifications of the system have first to be clearly determined. How are the logical connections between biological processes and molecules going to be represented? Will the system scale and work even with large biomedical input size? What is meant by biological knowledge? How does it fit currently available biomedical data? These questions will be addressed in Chapter 2, introducing description logics as a mathematical framework of choice to perform the MoA implementation later on. I propose an analogy considering living organisms as machines to describe a portion of the biomedical knowledge, relevant to drug discovery.

1.4.2.2 Chapter 3 - The Functional Therapeutic Chemical Classification System (Implementation)

The specification finds an implementation in Chapter 3, with the Functional Therapeutic Chemical Classification System (FTC). The classification features a representation of over 20,000 modes and mechanisms of action categories, inside which approved drugs have been automatically classified. The work done was evaluated against existing solutions and is publicly available via a web application. As expected, on average drugs are present in numerous MoA categories. This observation can be used to perform different types of analysis and generate drug repositioning hypotheses.

1.4.2.3 Chapter 4 - Systematic drug repositioning analysis

The content of the FTC is analysed in Chapter 4 from different perspectives. First, the relationship between the concepts of a drug's structure, function and indication is discussed in this chapter. Secondly a list of drug repositioning hypotheses is derived from this preliminary characterisation, covering a wide range of therapeutic areas. The hypotheses are used to analyse in a systematic fashion the drugs' off-label uses. Finally hypertension and Alzheimer's disease have been selected as use-cases in order to further investigate drug repositioning opportunities.

1.4.2.4 Chapter 5 - Outlook and future work

The outcomes of the thesis work are put into perspective: what part of the drug discovery process has been covered? What are the next logical steps for future work? Can the MoA representation be further improved and how? The pitfalls encountered during the work are discussed and I present my vision towards a simpler knowledge representation system for the biomedical domain. Chapter 5 also discusses alternative analyses that can be performed with the content of the FTC, in particular against gene expression data.

chapter 2

Description logics and biomedical knowledge (Specification)

Key points

- Organisms can be compared to complex black box machines, namely devices with a hidden and unknown internal functioning.
- This analogy helps to define the study, needs and representation of biomedical knowledge.
- Description logics (DLs), part of a mathematical framework developed to formalise concepts, can be used to study the molecular black box machine and query over recorded biological knowledge. DLs provide the means to represent terms and logically link biological modules.
- The framework presented integrates with current life-science information, such as biomedical ontologies (OBO) and databases, and is theoretically highly scalable ($\mathcal{EL}^{++}$ or OWL2 EL profile).
- The notions introduced and discussed set the groundwork for the representation of the *mode of action* and the implementation of a new resource built on these principles: the Functional Therapeutic Chemical Classification System (FTC).

Author's comment

This chapter is a summary and record of my experience with the Web Ontology Language (OWL2) and knowledge representation for the biomedical domain. I present a theoretical analysis of the representation of the mode of action and its scalable implementation. A software library to assist the development of programmatic solutions (Brain) is briefly presented in this chapter. I invite the reader to refer to the original publication (Croset et al., 2013a) for more detail if wanted.

2.1 Introduction

Science (from Latin *scientia*, meaning *knowledge*) is a *systematic enterprise that builds and organises knowledge in the form of testable explanations and predictions about the universe* (Wikipedia, 2014i). More specifically, biomedical sciences handle the subset of knowledge related to living organisms. Understanding the biological world is relevant to society as it provides, among other things, some valuable insight to treat and cure diseases. In order for our knowledge to grow, discoveries and evidence have to be recorded, structured and shared with the community and society. Traditionally, biological knowledge is preserved in a narrative fashion, inside textual documents, such as a journal article for example. However, natural language is ambiguous, informal, and impairs an efficient reuse of the information. More recently, with the advent of computer systems, some biological knowledge is stored in a structured way inside databases or ontologies (Brooksbank et al., 2014). Biological entities and concepts have identifiers, enhancing the dissemination and reuse of the information, as well as enabling an efficient integration of multiple datasets. Despite the improvements made towards a more meaningful and consistent representation, I argue that the *logic* in bio*logic* is still under-represented. Indeed, it would be valuable to formally derive and prove new facts from existing ones, the same way it is done in algebra for instance. This chapter presents an original approach towards this goal, using description logics (DLs) as a mathematical framework.

In this regard, I first illustrate how the study of organisms is analogous to the theoretical study of a black box machine. From this simple fundamental model

and its requirements, I then explain how DLs can, to some extent, capture the internal logic of the cellular machinery. Finally I discuss how this approach can be combined with existing biomedical data, in order to implement automated and scalable solutions. This theoretical work serves as the basis to formally define the concepts of mode and mechanism of action, central points in deriving drug repurposing hypotheses.

2.2 Biomedical knowledge

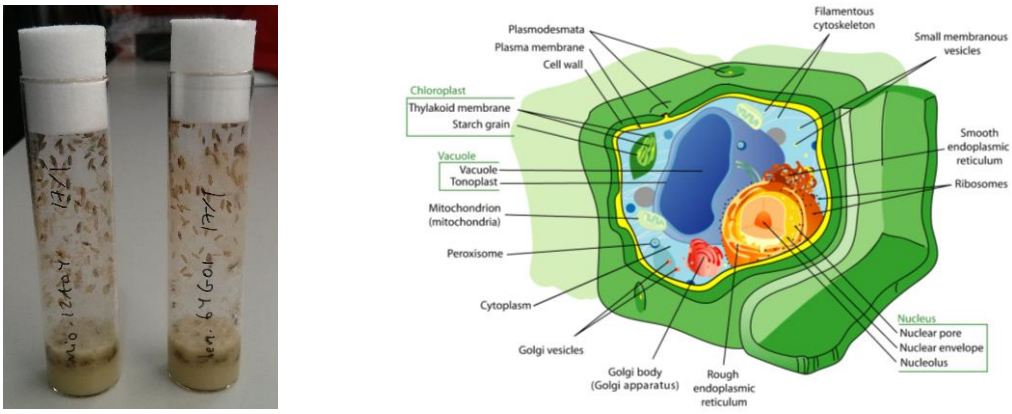
Life sciences addresses the study of living organisms. Just as in any other scientific discipline, biological researchers first collect data and evidence, which will then be turned into knowledge based on human interpretation (Antezana et al., 2009). In order to be efficiently reused, understood and shared with the scientific community, some of this knowledge can be represented by a process known as *formalisation*.

2.2.1 Contemporary formalism in biomedical sciences

Formalism can be described as *an abstraction process by the human brain in order to model a system in mathematical terms*. Formalising a problem helps to better analyse its complexity. Mathematical models (e.g. equations) are powerful, as they can eventually predict outcomes, once the system is understood well enough. Different scientific disciplines employ different formalisms, depending on the problem studied and research questions being asked.

2.2.1.1 Formalism varies among natural sciences

When I was a college student, I always found biology to differ from other natural sciences such as chemistry or physics for instance. The latter disciplines were built around a particularly strong mathematical scaffold; fundamental phenomena, like the speed of a body, were concisely captured by a function such as $v = d/t$. Chemical reactions were simplified in the form of an equation such as $H_2O = HO^- + H^+$ for instance. Despite being approximations of natural processes, these models and equations were useful to learn the discipline. Moreover, once the core concepts were understood, some supplementary layers of complexity were logically added on the top of the known things: static representations of chemical



(A) (B)

Figure 2.1: Examples of capture of information in biology. (A) A flask containing hundreds of individuals of the species *Drosophila melanogaster*. The population is identified with a label on the container. (B) Schema of a canonical plant cell, similar to the one found in textbooks. Parts of interest are annotated with terms, following a descriptive approach. Pictures from Wikipedia and courtesy of Sarah H Carl.

reactions were transformed into dynamic ones, or it was possible to derive the trajectory of a projectile from its speed and direction, for example.

The formal representation of a system is important, as it clarifies the meaning of the concepts of interest for the community. For instance, speed has a very explicit and clear definition captured by its equation, which can be unambiguously understood and interpreted accordingly. Finally, the most important feature of a formal system is, I believe, that it enables predictions to be made. It becomes possible to infer behaviours and results on the sole basis of the theory. Complex systems can be built and fully understood, relying solely on fundamental principles.

The study of biomedical sciences was not guided by such strong mathematics. Core concepts such as *evolution* or *gene* were mostly described in a textual way, a sentence usually defining the meaning; it was impossible to combine concepts in order to create new ones. As biological organisms obey the law of physics and chemistry, one would therefore assume that these frameworks could assist in the study of the living world on their own. They do to some extent; it is for instance possible to represent and model the series of chemical reactions related to a biological pathway using standard chemistry (Le Novere et al., 2006). However, biomedical sciences present particular challenges that cannot adequately be

solved by the traditional molecular formalism. First of all, biological bodies are extremely complex from a chemical perspective, either in terms of size or types of compound present (3 x 10^{27} atoms in the adult body and a minimum of 25 atom types) (Nielsen, 1999). Secondly, some high level phenomena have an unknown molecular basis, and it becomes therefore impossible to capture formally the system while solely relying on chemistry. Phenotypes and diseases are good examples of this category.

Because of these obstacles among other things, biomedical sciences are traditionally less formalised than their counterparts, the biological knowledge being often captured by a textual description inside a document, or sometimes with the help of a conceptual schema (Lazebnik, 2002). Figure 2.1 and 2.2 present examples of media used to convey, record and communicate life sciences knowledge.

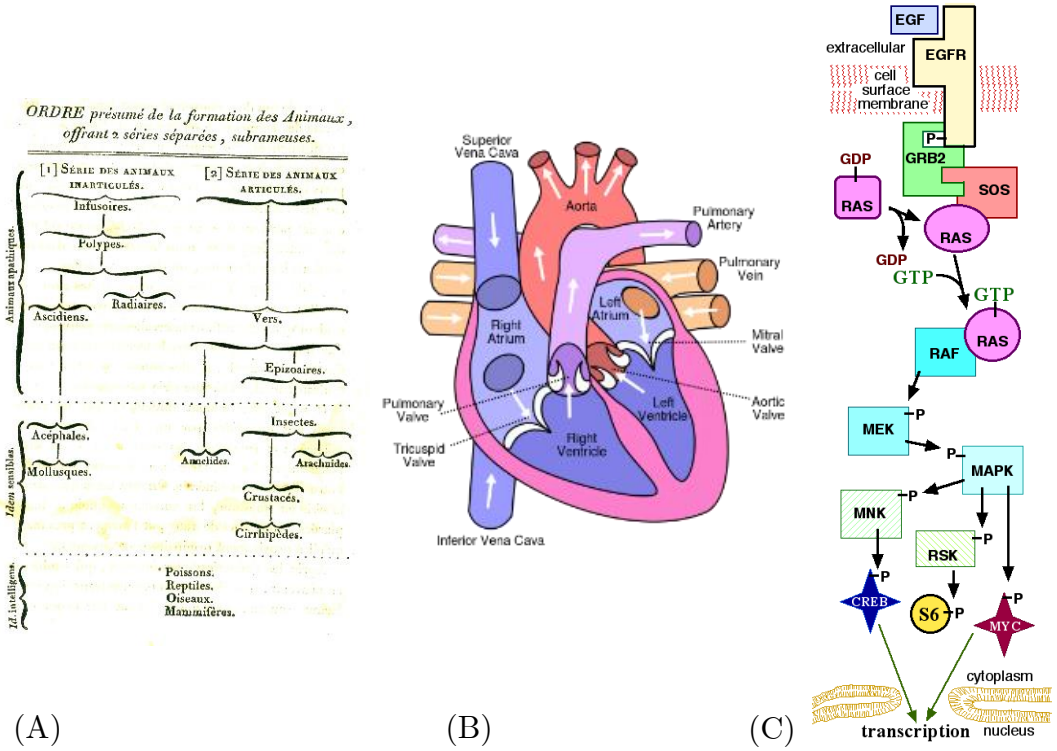


Figure 2.2: Formalisation in biomedical sciences. (A) Diagram showing the evolutionary taxonomy of invertebrates. Drawn by Jean-Baptiste Lamarck's in 1815. (B) Mechanistic illustration of the cardiovascular system. Arrows illustrate the flow of blood and the logical connection between the annotated parts. (C) MAPK/ERKsignalling pathway. Schema of the cascade of molecular events leading to activation of transcription factors. The logic of the system is informally captured using arrows, colours and shapes. Images from Wikipedia.

Nowadays, with the advent of computers and the Internet, known biological entities and concepts are further stored inside public databases. Often, a manual curation step over the published literature improves the consistency and correctness of the data (Brooksbank et al., 2014). The structured information makes it easier for the community to retrieve the data and to perform statistical analyses on it. Nonetheless, this framework is not fully formalised; for example it is not possible to mathematically prove why a drug could be useful for a disease, such as one can logically derive with a series of steps the value of the variable x out of the following equation: $2x = 4 + x$.

How can one further formalise biomedical knowledge to assist the development of new medicines and the study of the living world? A naive approach would try to simplify the system life scientists are working with into a meaningful analogy (Lazebnik, 2002). In this regard, I will present how the study of a living organism can be compared to the study of a black box machine, namely a device for which nothing about the internal workings is known (see Figure 2.3).

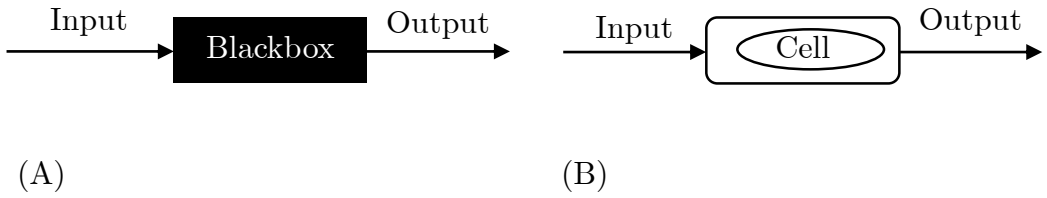


Figure 2.3: Black box model. (A) Schematic representation of a black box. Given an input, an observable output is produced. The internal workings are supposedly not understood. (B) Cells or organism are black boxes: they can carry various and observable functions from an input, yet the internal workings are not necessarily fully understood.

2.2.1.2 Organisms as complex machines

From the perspective of drug discovery, an organism (word related to *organisation*) can be broadly simplified as *an assembly of molecules functioning as a stable whole*. This characteristic makes organisms similar to a machine in the generic sense, which can be described as *an assembly of parts functioning to meet a particular goal*; in the case of the organism, the intrinsic goal is survival or reproduction. Based on these definitions, biomedical sciences can be seen as

the sciences of preventing and repairing dysfunctioning organisms - or molecular machineries.

According to this analogy, the study of a living organism can therefore be compared to the study of a complex black box machine, composed of a large number of physical parts carrying a certain number of internal functions and acting together in an organised fashion.

Now assuming that a certain community has access to such a machine and wants to study it for various reasons, how can it theoretically be done? The analogy becomes insightful at this stage, as there are plenty of complex man-made devices the reader can relate to, and which will serve to illustrate the thought process. I will consider an airplane as example, because it is a complicated device with a straightforward goal: safely flying in the air. A similar exercise can be done with a radio (Lazebnik, 2002).

Supposing that a fully functional airplane was found somewhere and the only thing known about it was that the device is capable of flying. The aim is to understand as accurately as possible how the machine works, in order to have enough knowledge to be able to fix it in case it gets broken or malfunctions in the future. In order to address this problem, I and Lazebnik (2002) argue in favour of a straightforward descriptive approach, in line with the way biological sciences are performed, namely describing the device in as much detail as possible.

The first task done would be to schematise the device, as shown in Figure 2.4-A. Then the fundamental physical parts would get annotated with arbitrary names (Figure 2.4-B). With an increased understanding of the physical modules composing the airplane, it would then be possible to discover the roles played by the various parts of the machinery. Objects would receive functional annotations, namely an explanation of what they do in the overall flying process, as shown on Figure 2.4-C. Up to this step, the system would be characterised as a collection of discrete physical and functional modules, each isolated from one another. Finally, in order to appreciate the machine as a whole, it would become mandatory to link the modules based on their relation types. Figure 2.4-D shows the high level logical organisation of the machine, which integrates the parts to understand the overall process.

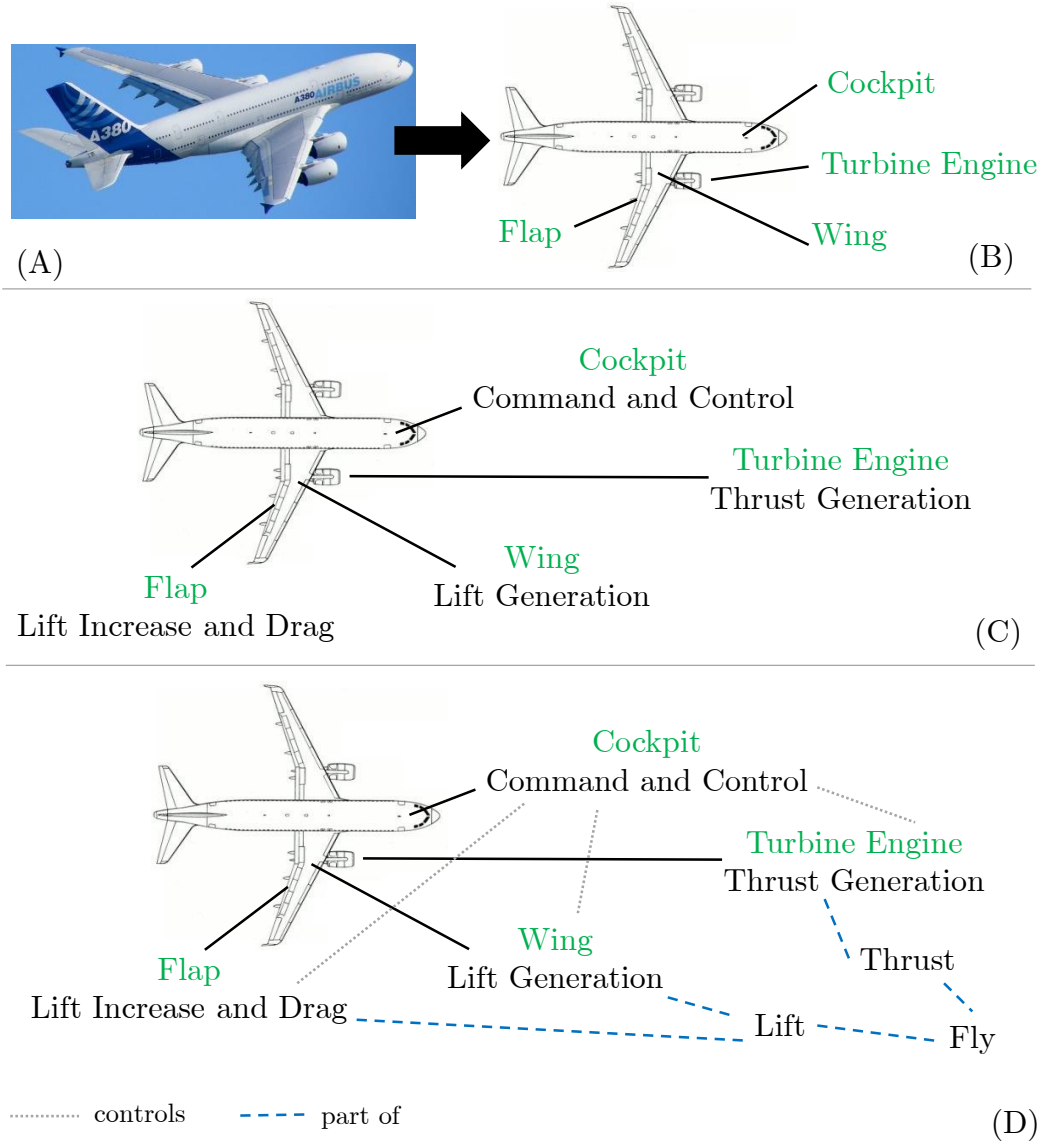


Figure 2.4: Study of an airplane using a descriptive approach. (A) The device is first schematised. (B) Physical parts are annotated with terms and concepts. (C) Physical parts receive functional annotations. (D) The different modules are linked based on their relative roles.

Unveiling the internal logic of the black box machine using this descriptive methodology would later allow extensive querying of the model of the system and simulation of its behaviour. For instance, assuming a problem was identified with the lift of an airplane, preventing it from flying. From the model, it would be possible to retrieve all the known parts directly or indirectly involved in the lift process, derive a list of potentially faulty components and suggest ideas on how

to logically repair it and restore its ability to fulfil its primary goal, flying.

The black box machine analogy allows one to better understand the requirements of a formal descriptive framework in order to capture biomedical knowledge. From a drug discovery point of view, organisms can be fundamentally reduced to machines. The descriptive process of studying such machines I presented is analogous to the approach researchers use to study organisms, understand diseases and find treatments for them.

In order to move from an informal characterisation into a defined framework, it is first necessary to determine what is required for a descriptive approach to successfully capture biology. The next section will identify some of the concrete needs for biomedical knowledge, derived from the theoretical model of the molecular black box machine and its study.

2.2.2 Requirements for biomedical knowledge formalisation

In order to be efficiently reused and shared, biomedical knowledge has to be formalised. One way to investigate the living world consists of considering organisms as complex machines; descriptions and annotations of the parts of the machine then help to represent the knowledge and understand better the functioning of the whole device. However, in order to be meaningful, this descriptive approach needs to fulfil a series of criteria introduced in the following sections.

2.2.2.1 Mathematical framework

Extending a mathematical field is a key feature of any formal framework. Mathematics help to accurately formulate problems and provide the generic means to solve them. In the case of biomedical knowledge, the ultimate aim is to reduce the burden of diseases for society. To achieve this goal, a formal framework must first be able to capture biological information. Secondly and more importantly, it should be possible to further exploit the framework to deduce and prove assumptions over it. For instance, the mathematical branch of geometry handles questions related to shapes and space. Even if this framework provides only a simplified view of reality, geometry provides the formal means to attest to the validity of a building's blueprints or optimise land exploitation. Similarly, in

the biomedical domain, one should expect to be able to deduce implicit facts or formulate new hypotheses regarding potential treatments directly out of the mathematical formalisation, which is currently not the case. Connecting descriptive biomedical knowledge with mathematics also insures that the framework can benefit from the latest progress in the field. It helps different communities to work together on the same issues from different angles. Formulating biomedical knowledge in mathematical terms also opens the door for computer sciences to assist later on with the implementation of a digital solution.

2.2.2.2 Definitions

The annotation of the parts of the machine requires the usage of a new and specific vocabulary. Words and concepts can be ambiguous, especially when the system analysed is complex such as a living body. A recent illustration of the importance of definitions in biology is the debate over the ENCODE project conclusions (Editorial, 2013); different scientists have different interpretations of the word *function*, which shift the explanation of the results. Semantics, the investigation of the meaning of symbols and words, can assist in this task and help to specify the intended meaning of a concept. Moreover, a formal framework for biomedical knowledge should be able to define any type of things: real-life objects, such as *molecule*, *protein* or *cell* for example. Abstract biological processes like *blood coagulation* or diseases such as *cancer* should also be part of the framework, as they are essential concepts in biomedicine. Finally, in order to appreciate the logic of the machine as a whole, it is mandatory to be able to link concepts and words, in order to show how parts of the machinery interact together. The meaning of such relations should be explicit and unambiguous, just as the definitions of concepts.

2.2.2.3 Hierarchies and abstraction

In practice, organisms are probably more analogous to Rube-Goldberg machines than airplanes (see Figure 2.5), with an internal logic sometimes difficult to understand on its own; this results in practice in a tangled network of chemical wiring, which can be abstracted and simplified into functional modules (Hartwell et al., 1999) (Ravasz et al., 2002) (Machado et al., 2011) (Fisher and Henzinger, 2007). Biomedical knowledge has to deal with entities ranging from chemical drugs to

high level concepts such as *species* or *biological processes*. All these layers have to be integrated and linked in order to understand the machine as a whole. In this regard, it is critical for the formal framework to support abstraction and enable the representation of hierarchical information.

Taxonomies have always been at the heart of biological sciences; take for instance the work of Carl Linnaeus and the Systema Naturae (von Linné and Lange, 1770). Classifications are further used to organize species, protein and chemical families, to name a few examples. Historically speaking, categorical information has provided a good and intuitive framework to capture biomedical knowledge; therefore, any attempt for further formalisation must be able to handle this type of data, as well as to leverage its use.

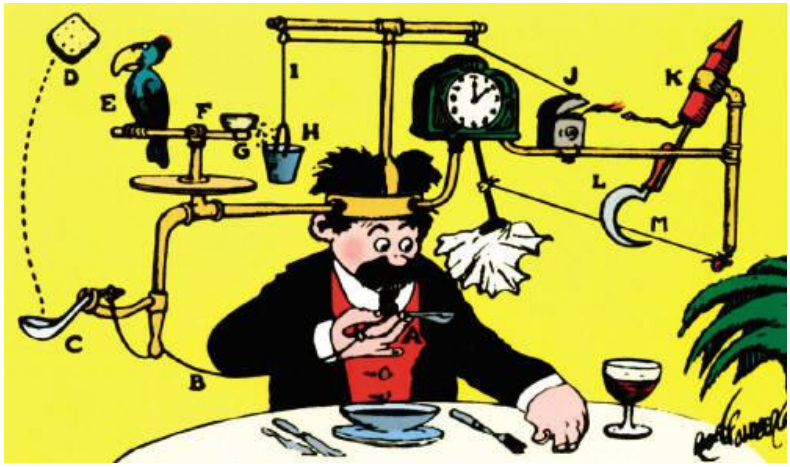


Figure 2.5: Rube Goldberg machine: over-engineered machine that performs a very simple task in a very complex fashion, usually including a chain reaction (Wikipedia, 2014h). The picture shows the "Self-Operating Napkin". When the spoon soup is raised, a cascade of events are triggered ending as the napkin coming toward the man's face. The task performed is relatively trivial, yet many steps are needed to execute it. Organisms are assumed to be analogue to Rube Goldberg machines because of evolution; the internal wiring is not necessarily straightforward and progressively evolved and changed (Ravasz et al., 2002). Illustration from Wikipedia.

2.2.2.4 Distributed and scalable

Studying a system as elaborate as an organism requires the collaboration of a large number of persons, working in parallel on different facets of the problem. All these individuals must be able to access and share their knowledge, in order to

communicate and be aware of the latest progress. Disseminating information has been usually performed via printed literature, which is being replaced at the time of writing by the World Wide Web and the Internet. This shift of infrastructure allows the processing of more and more information in a digital context, where computers can perform an increasing amount of the work in an automated fashion. This characteristic is particularly interesting for biomedical information, as it is possible to use computers as part of the formalisation process. However, this approach leads to challenges, in particular scaling issues. As organisms are of high complexity, it is therefore important to opt for a formal solution able to cope with problems of large input size. It is assumed that biomedical knowledge will only grow bigger with time, so a mathematical framework should take care of this concern, in order to be future-proof. Finally, it is supposed that biomedical knowledge is and always will be incomplete; in order to be powerful enough, the mathematical formalisation has to able to handle missing information.

2.2.2.5 Molecular dynamism

One of the characteristics of living forms is their dynamism; as organisms are made of molecules, they are subject to the laws of chemistry and molecular dynamics. Organisms can be defined in chemical terms as *semi open systems* (Meng et al., 2004), which emphasises a strong relationship with the environment. Formalisms coming from chemistry, such as conservation of mass (Wikipedia, 2014b) or thermodynamics (Wikipedia, 2014j) can represent and solve such systems, yet they are not suited to handle more abstract concepts from the biomedical domain. An ideal formal framework for biomedical knowledge would appreciate the impact and interaction with the environment, yet this requirement is extremely challenging in regards to the number of chemical reactions to be considered (Meng et al., 2004). Moreover, the chemical formalisation strongly relies on kinetic parameters to capture behaviour, which makes them vulnerable to missing knowledge. Finally, another concern is the effect of chemical concentration in regards to the function. For instance, the action of a drug strongly depends on its administered dosage. Understanding in detail the biological machinery and deriving correct predictions out of it implies considering molecular dynamics.

Formalising biomedical sciences can be done using a descriptive methodology; this approach has been used since the origin of the field and is an intuitive way

to represent biological systems and facts. I presented in this section the theoretical requirements deriving from the descriptive methodology and more generally biomedical sciences. The coming section illustrates how DLs can address some of these requirements in order to mathematically formulate and further use biomedical knowledge.

2.3 Description logics for biomedical knowledge representation

Description logics (DLs) are part of a family of formal languages used to represent the knowledge of a domain of interest. The plural form of the word *logic* indicates that a multitude of languages exist, each one of them characterised by a certain type of expressivity, as it will be seen later in this chapter.

Briefly summarised, the framework provides means to group and categorise sets of things based on shared properties, and allows hierarchical classifications to be built. More complex logical structures, such as roles, are further available in order to enrich the modelling. While using DLs, the formalisation results in a graph structure, which evolved from *semantic networks* (Allen and Frisch, 1982). This characteristic is important for the biomedical domain, as it matches ongoing graph and network representations, as it will be discussed in the coming sections.

DLs are well characterised from a theoretical point of view and implemented in the Web Ontology Language (OWL), a standard supported by the World Wide Web Consortium (W3C). For these reasons and because they address most of the requirements presented before, DLs are an ideal mathematical framework that can be used to formalise some biomedical knowledge.

2.3.1 Problems addressed by description logics

According to Gruber, DLs help to *specify a conceptualisation* (Gruber et al., 2009). In the case of life sciences, the conceptualisation is the molecular machinery, finding its specification being the task of the researcher. In this regard, DLs come with a series of tools to define words and concepts, and in particular their associated meaning or semantic. This feature allows terminologies to be built to describe the molecular black box machine. For instance, it is possible to capture

the relative difference in meaning between these three following abstract biological concepts using DLs: *positive regulation*, *negative regulation* and *regulation*. This task appears trivial for humans, yet the main motivation behind DLs is to be able to express such things in an unambiguous and formal manner, in order to deduce less evident information later on. With DLs, the interpretation of the meaning of a concept comes from its relation to other concepts. For example, the concept *mammal* is more generic than the concept *human*; this is asserted in DLs by stating that every instance of human is also a mammal. The mathematical interpretation is that the set of humans is a subset of mammals. An *is a* relation could also be drawn between the two concepts if they were represented as nodes (see Figure 2.6). This feature enables DLs to unambiguously define the meaning of words from a mathematical perspective. It is also possible to define in a similar way the meaning of relations and to use them to represent the logic and wiring of the molecular machine later on. Taken together, these theoretical properties alone address the needs for a mathematical framework (section 2.2.2.1), capable of handling definitions (section 2.2.2.2) and abstraction (section 2.2.2.3).

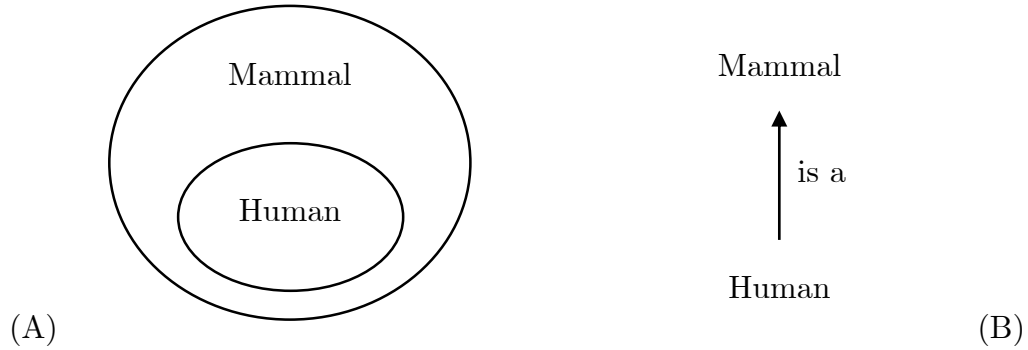


Figure 2.6: Example of problem addressed by description logics. (A) The concept *Human* is more specific than the concept *Mammal* (all humans are mammals), which can be represented as embedded sets. (B) Same logic captured, representing the concepts as nodes and the relation as edge. The mathematical meaning of (A) (sets of instances) is more accurate than the meaning of (B), yet both representation can exist in practice, in particular in biology.

The relations between entities and concepts are meant to formally express the known truth about a domain of interest (Stevens et al., 2007) (Krötzsch, 2012) (Hitzler et al., 2009). This type of construct is called an *axiom*. A set of axioms constitute a *knowledge base* (also called *ontology*). The main interest of a knowl-

edge base is the possibility to use the meaning of the axioms to deduce implicit things. Consider as an example the question, "What are the different types of mammals?". From the example before, *Human* would be logically deduced from the connection between the two categories (Figure 2.6).

From the formal representation it becomes possible to generate proofs and ask queries over the knowledge base; this operation is called *reasoning* and consists of two tasks: subsumption and consistency checking. Briefly the structure of the knowledge base arises from subsumption; concepts get classified into a taxonomy based on the axioms. Consistency checking can detect inconsistencies or contradictions that might be present in the knowledge base and report them. These services will be presented in details later in this chapter. It is important at this point to understand that the axioms of the knowledge base can be used to either generate more knowledge or to assess the validity of the current representation in a formal fashion. Reasoning tasks can be performed by humans and, more interestingly, by computers too. In fact, DLs have been developed for this very reason, to render domain knowledge computer understandable. The tight connection between computers and DLs is adequate for biomedical sciences, where it is expected to face large-scale data, beyond a sole human brain's capability to handle at once. The interoperability with computers makes DLs adequate to address the requirement for a distributed and scalable framework (section 2.2.2.4).

DLs are efficient at representing terminologies and logical relations, yet they are unfortunately not appropriate to represent dynamic and temporal knowledge (Kim et al., 2008). This framework presents limits in regards to the molecular dynamicity requirement (section 2.2.2.5) and cannot handle well this facet of the machine. Work-around solutions will be presented later to the reader (Chapter 3).

Because DLs appear to address well a majority of the requirements for biomedical knowledge formalisation, I believe the framework to be suitable for the study of the molecular machinery. DLs can help to annotate and connect the logical parts in order to create a knowledge base, useful to understand the overall functioning of the system. Reasoning services allow for the querying and further use of the knowledge in a formal fashion, as expected from any mathematical framework. Finally the computer implementation of DLs is well studied, therefore part of the work can be automated, in order to come up with a scalable solution. Scalability guarantees the long-term success of the framework, but comes at the

cost of expressivity.

2.3.2 Expressivity and complexity

Wikipedia defines expressivity as *the breadth of ideas that can be represented and communicated*. Consider for instance a molecule of methane; this concept can be represented in at least three different ways, more or less expressive and detailed, as shown in Figure 2.7.

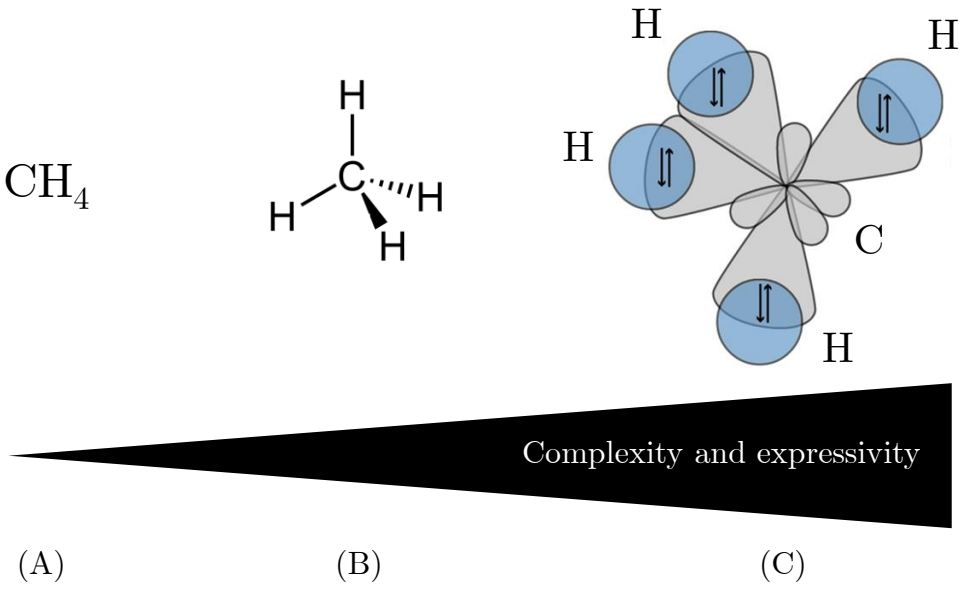


Figure 2.7: Complexity and expressivity. Increasingly expressive and complex representation of a methane molecule. (A) Molecular formula. (B) Representation of the chemical bounds and their types between the atoms. (C) Representation using molecular orbitals. Images from Wikipedia.

What abstraction level is the correct one? It depends on the type of questions being asked on the model; all these three representations are legitimate and useful in practice for different purposes. The main difference between these models is their expressivity. Model 2.7-A for instance is less expressive than model 2.7-C; it conveys less detailed information, yet it is easier to understand and maybe more suited to study different types of problems. DLs are a family of logics, and just as with the methane molecule and depending on the type of axioms considered, they can model concepts in a more or less expressive fashion. As a rule of thumb, the more expressive the language is, the more complex it is. In the

case of DLs, the complexity refers to the computational complexity, namely *the time taken for the program to finish performing reasoning* (Krötzsch, 2012). The more axiom types are available to the user, the more time it will take to deduce their entailments. Complex problems are to be avoided in computer science as it will take more time to return a deduction from the content of the knowledge base, if a deduction is returned at all. This limitation comes from the hardware and the way computers are currently built (Turing, 1950) (Neumann and Burks, 1966). In this respect, a scalable solution is a trade-off between expressivity and complexity. One would like to be as expressive as possible in order to study the molecular machine, yet able to deal with the very large input size of biomedical information. Fortunately, one of the strength of DLs is that their computational complexity has been particularly well studied (W3C, 2014b) (ter Horst, 2005). It is possible to work with a restricted set of axiom types, guaranteeing an acceptable complexity, appropriate for large-scale implementation. These fragments of DLs are also called profiles or subsets; one of them, the $\mathcal{EL}^{++}$ profile, is particularly interesting for biomedical sciences (Baader et al., 2005) (Baader et al., 2008) (Hoehndorf et al., 2011a). I will focus in the rest of this chapter on this very fragment, as it is possible to reason over an $\mathcal{EL}^{++}$ knowledge base in polynomial time. This characteristic makes reasoning tasks a *tractable* or so called *easy problem* (Cobham, 1965), which ensures that the framework can still work and scale for extremely large datasets. Despite offering a limited expressivity, the $\mathcal{EL}^{++}$ profile provides the means to address a good portion of the requirements for biomedical knowledge formalisation, as will be presented in the coming section.

In the quest to formalising biological information, DLs have the very clear advantage of a well characterised computational complexity, ensuring the creation of realistic practical solutions with the help of a computer, and not theoretically limited by the size of the data. On the contrary, the computational complexity of simulating and modelling biological systems at the level of the chemical reaction is less neatly defined (Meng et al., 2004) (Gillespie, 2007) and appears much more complex and challenging; the fuzziness around the hardness of this task leaves an unanswered question as to whether modelling based only on chemical formalism will scale to systems as large as a cell.

2.3.3 DLs' components and relation to life sciences

Because of scalability concerns, it is wise to stay within the boundaries of relatively low complexity (section 2.3.2). To address that matter, I have presented the $\mathcal{EL}^{++}$ profile as good candidate language, combining a decent expressivity for biomedical knowledge yet featuring constructs of an acceptable computational complexity. I will now present these constructs and components and explain how they relate to the life science domain. DLs are a theoretical framework and find a computer implementation as the Web Ontology Language (OWL2) (W3C, 2014c). The $\mathcal{EL}^{++}$ family is more specifically implemented in the OWL2 EL profile (discussed later). Despite being very close conceptually, these two frameworks have a few differences on the terminological level; I will therefore also mention the OWL2 equivalent names and symbols as a reference. OWL2 will be briefly discussed later in this chapter.

2.3.3.1 Description logics entities

In order to represent the domain knowledge, DLs offer three kind of core entities, or building blocks: named individuals, roles and classes. These entities are used to represent the world, and in this case, the components of the molecular machine. The first question that comes to mind is why these three types? The choice is purely arbitrary and likely derives from ancient Greece and the early work on categorisation done by philosophers such as Parmenides and Aristotle (Wikipedia, 2014d). It appears that these three types can represent quite a lot of information, and they are fairly intuitive for humans to understand and reason over. They are an acceptable way to abstract the world around us.

Named Individuals

OWL2 terminology: individuals

The first building block handles individuals. Individuals refer to real-life instances and objects: this squirrel in a tree, this single molecule of water or glucose, this pen on a desk are all examples of individuals. Individuals are at the centre of DLs modelling. Everything else gravitates around them; all the further representation is done in regards to them. Every object can be considered as an individual.

Surprisingly, individuals are very rarely represented in life sciences. For example when one speaks about a particular protein, the reference is made towards the canonical version of the protein and not to the very one instance mixed with the millions of identical others. The same applies for diseases and species; the life scientist is concerned with extracting generic patterns and categories and reasoning at a more abstract level. This can be achieved by considering sets of individuals, so called *classes* or *concepts*.

Classes or concepts

OWL2 terminology: classes

The second building block type handles concepts and terminologies. DLs concepts are interpreted as mathematical sets, namely groups of objects or individuals. Concepts represent an abstraction over instances and fit biological reasoning well. For example, the concept human contains at least two individuals, the reader and myself. Most biomedical ideas can be described as a concept, including not only material entities such as molecules (e.g, *P53* or *paracetamol*) but also immaterial processes or functions (e.g, *blood coagulation* or *catalytic activity*). This characteristic makes DLs concepts very suitable to describe the molecular machine. The difference between a DL concept and its members is illustrated in figure 2-8-A.

N.B: DLs concepts are also sometimes called *classes*. Because the word *class* is less ambiguous than the word *concept* in the context of this work, I decided to use refer to DLs concepts as *classes* in the rest of the manuscript. The words *category* or *type* will also be used to refer to *classes*, later in the document. The reader can simply consider that *class*, *category*, *type* and *concept* refer to same idea in this manuscript.

Roles

OWL2 terminology: properties

The last building block deals with roles, namely how individuals relate one to another. Roles are more subtle and flexible to use than the two previous entity types. Basically they exist to capture a logical link between two individuals, and are so called binary in this respect (see Figure 2-8-B). In the study of the molecular black box machine, roles connect the modules and processes; they can represent the logical connections of the machine. Examples of roles commonly

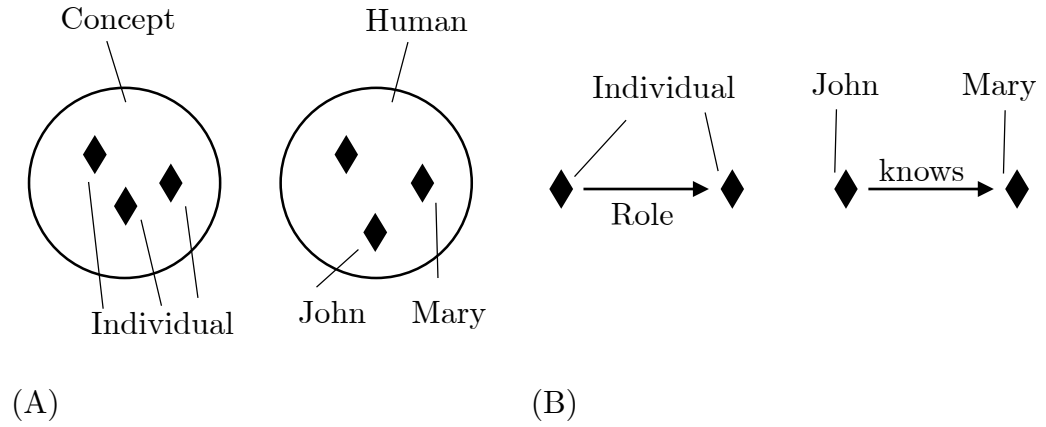


Figure 2.8: Description logics core entities: (A) Concepts and Individuals: a concept or class is a set containing some individuals. On the example shown, Human is a concept, John and Mary are both distinct individuals. (B) Roles and Individuals: role are linking two individuals. In the example, John and Mary are still individuals linked by the role *knows*, specifying their relationship.

used in the biomedical domain are: *regulates*, *part-of*, *involved-in*, etc. Note that roles are only linking individuals; they cannot be used to directly link classes.

2.3.3.2 Axioms

In DLs, the axioms are the reflection of our current vision of the world; they express the mathematical truth and help both humans and machines to interpret the meaning conveyed by the entities of the knowledge base (classes, individuals and roles). Axioms enable deductions to be made. For instance, let's assume that there is an enzyme called *thrombin* which is somehow involved in *blood coagulation* (first axiom) and that a compound named *ximelagatran* affects the activity of *thrombin* (second axiom). From these axioms, one could conclude that the *ximelagatran* might logically affect *blood coagulation*. The progression from a set of axioms to a conclusion is called reasoning, and, as discussed previously, this operation can be performed by humans or by computers with the help of a program, called a reasoner. Different types of axioms exist, depending on what needs to be modelled.

Assertional axioms (ABox)

The assertional axioms (or assertional box - ABox) are used to assert the truth

about individuals. Although actual individuals are rarely represented as such in life sciences, it is still important to understand these fundamental types of axioms before looking at more complex ones.

Concept assertion

OWL2 terminology: class assertion (or types). Concept assertion axioms link individuals to their classes or types. For example, let's consider a knowledge base containing one named individual called *John* and one class named *Human*. The axiom asserting that John is a human is written *Human(john)* or sometimes *john : Human*. It states that john is a member or instance of the class *Human*.

Role assertion

OWL2 terminology: property assertion (or facts). This axiom captures the relationship between two individuals connected via a role. For example, consider two named individuals, John and Mary, and a role, *knows*. Asserting the fact that John knows Mary as shown in figure 2-8-B is written knows(john, mary) or (john, mary) : knows in DLs. Role assertions are sometimes said to create triples or sentences; this axiom type helps to render logical networks or graphs.

Terminological axioms (TBox)

The terminological box contains the axioms related to classes, which are therefore of primary interest for the biomedical domain. TBox axioms formalise the relationship among classes in regards to the individuals they contain.

Concept inclusion ($\sqsubseteq$)

OWL2 terminology: subClassOf. Concept inclusion expresses the relationship between two classes, one being more specific than or subsumed by the other. Assuming that there are two classes, *Human* and *Mammal*, the concept inclusion axiom $Human \sqsubseteq Mammal$ entails that all instances of *Human* are also instances of *Mammal*, or in other words all humans are mammals. Note that even if the assertion was based on a mental reasoning about individuals, in practice we do not see any individual names, only classes. The axiom can be visualised in Figure 2-9, alongside concept assertions.

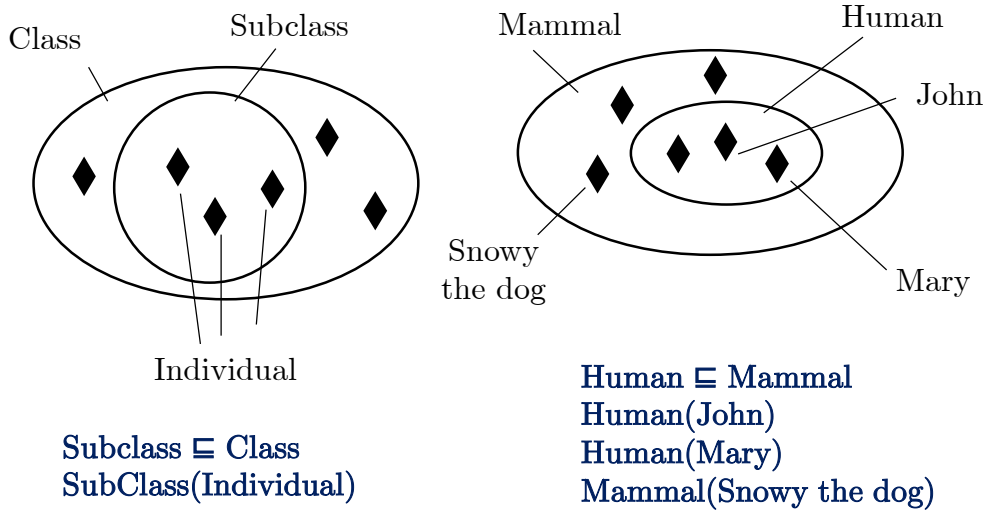


Figure 2.9: Examples of description logics axioms, in blue with their graphical representation (concept assertions and inclusion axioms). Axioms specify the semantics linking the basic entities (individuals, classes and roles). From a series of axioms or knowledge base, it is possible to deduce information. The question "What are the mammals present in the knowledge base?" would return as a result Snowy, but also John and Mary, even if they are not directly declared as such, from the semantics encoded in the axioms.

Concept equivalence ($\equiv$)

OWL2 terminology: equivalentTo. It is sometimes interesting to define a set of individuals as equal to another set of individuals. This construct is mostly used to create new classes from existing ones or to query a knowledge base. This axiom can state for example that the class *Human* is equivalent to another class or combination of classes, as will be seen later in the chapter.

Relational axioms (RBox)

The meaning of a role can be further specified in regards to the other roles of the knowledge base, in a similar way as is achieved with terminological axioms.

Role inclusion ($\sqsubseteq$)

OWL2 terminology: subPropertyOf. A role inclusion axiom defines the connection between two roles. Assuming that *R1* and *R2* are both roles in a knowledge base, and that the following role inclusion axiom holds: $R1 \sqsubseteq R2$, it means that all the time a pair of individuals is linked by the role *R1*, this pair is also linked

by *R2*. The role inclusion axiom is analogous to the concept inclusion axiom for roles. A biological example concerns regulation: the role positively-regulates is subsumed by the role *regulates*: *positively-regulates* $\sqsubseteq$ *regulates*.

2.3.3.3 Constructors

It is possible to specify the semantics of the core entities using axioms. The expressivity seen so far is however fairly limited. Fortunately the DLs constructors provide new means of expression as illustrated in the coming sections. Constructors allow for the composition of complex types from simpler ones analogously to how symbols of cuneiform scripts appeared to be used in some situations (see figure 2-10-A). Just as with axioms and entities, there are different types of constructors.

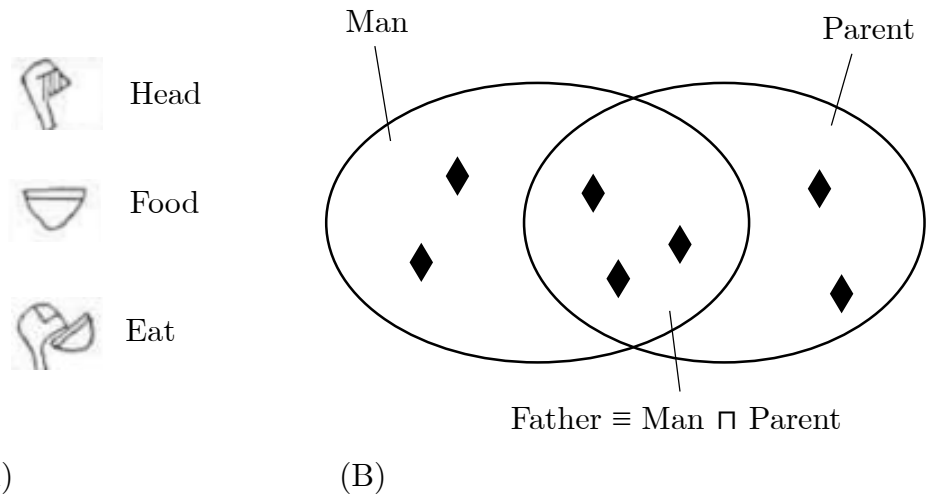


Figure 2.10: Description logics constructors. (A) Constructors help to compose with classes. The logic behind constructors is analogous to some of the semantic found in cuneiform scripts (3000 BC) (personal interpretation). (B) Concept intersection: the individuals member in the same time of the concept *Man* and *Parent* are asserted to be of type *Father* on the axiom presented.

Intersection ($\sqcap$)

OWL2 terminology: intersection (and, that)

The intersection constructor corresponds to the basic Boolean and operator used to describe overlapping sets. The figure 2-10-B depicts the meaning of the

construct. An intersection can be used to formulate new expressions, which are used to create more complicated axioms. The class *Father* can for example be expressed as being equivalent (equivalent concept axiom) to the intersection of the class *Man* and *Parent*. Both conditions have to be true in order to satisfy this expression. From this axiom, assuming one knows that an individual is both a parent and a man at the same time, one could deduce that this individual is also a father. Note that because of the equivalence axiom, the class *Father* is also inferred as being a subclass of *Man* and *Parent*; indeed if an individual is a father, it is therefore a man and a parent too from the definition.

Existential Restriction ($\exists$)
OWL2 terminology: existential restriction (some)

Existential restrictions encode a semantic subtle to grasp at first encounter (personal teaching experience). Yet they are particularly suited for the biomedical domain, as a means to link classes via roles, something not doable with the constructs introduced previously. An existential restriction captures the fact that all individuals of a type X are necessarily linked by a property to some of the individuals of a type Y. For example, the expression $\exists$ *part-of.Cell* refers to the sets of individuals that are necessarily part of a cell (see Figure 2-11-A). The presence of one of these individuals implies that here exists an instance of cell inside which they are located, no matter what. This expression can be combined with a concept inclusion axiom; for example, the class *Nucleus* can be expressed as follows: *Nucleus* $\sqsubseteq \exists$ *part-of.Cell*. The axiom entails that a nucleus is something that is always part of a cell. But only being part of a cell ($\exists$ *part-of.Cell*) does not necessarily qualify something to be a nucleus. Moreover, the axiom does not entail that all the cells have a nucleus. The example is represented and labelled in figure 2-11. Existential restrictions are very useful to represent biomedical knowledge; they allow one to link classes with roles, without explicitly referring to specific named individuals. Their entitlements are not too strong and often adequate for the biological world.

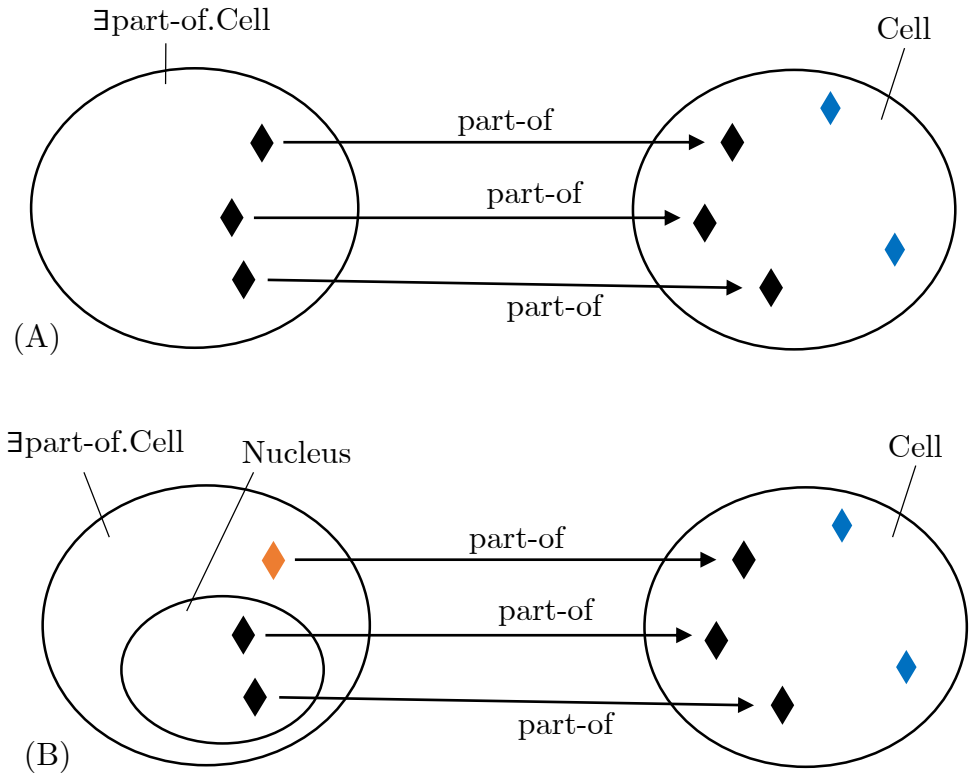


Figure 2.11: Description logics constructors: existential restriction (A) Example of existential restriction expression with the implication. Note that all individuals of ∃ *part-of.Cell* are linked to a cell individual via a *part-of* role. Yet some cell instances exist without being linked (in blue - for instance red blood cell). (B) Definition of the class *Nucleus* from the existential restriction construct: *Nucleus* ⊑ ∃ *part-of.Cell*, meaning that all instances of nucleus are necessarily linked to an instance of cell. There exist some instances of ∃ part-of.Cell not being nucleus instance (orange). This type of construct is commonly encountered in biology (e.g. Ashburner et al. (2000)).

Role composition (∘)

OWL2 terminology: chained properties (o)

The last constructor presented is the role composition. This construct enables chaining two roles together in order to create a new one. The typical scenario for this constructor is the *uncle* role, defined as a role inclusion axiom: *brother-of* ∘ *parent-of* ⊑ *uncle-of*. It is interpreted as if an individual X is the brother of a person with a kid, then X is the uncle of the child (see Figure 2-12-A). Closer to biology, an example of role composition is the expression *regulates* ∘ *part-of*, which can be used to create the inclusion axiom: *regulates* ∘ *part-of* ⊑ *regulates*.

This means that when the role *regulates* is followed by the role *part-of*, it could be simplified into a single *regulates* role. There is a special type of complex property inclusion called *transitivity*. It means that the role is chained with itself. For example: *part-of* $\circ$ *part-of* $\sqsubseteq$ *part-of*. The axiom is understood as if X is part of Y and Y part of Z then X is part of Z (see Figure 2-12-B).

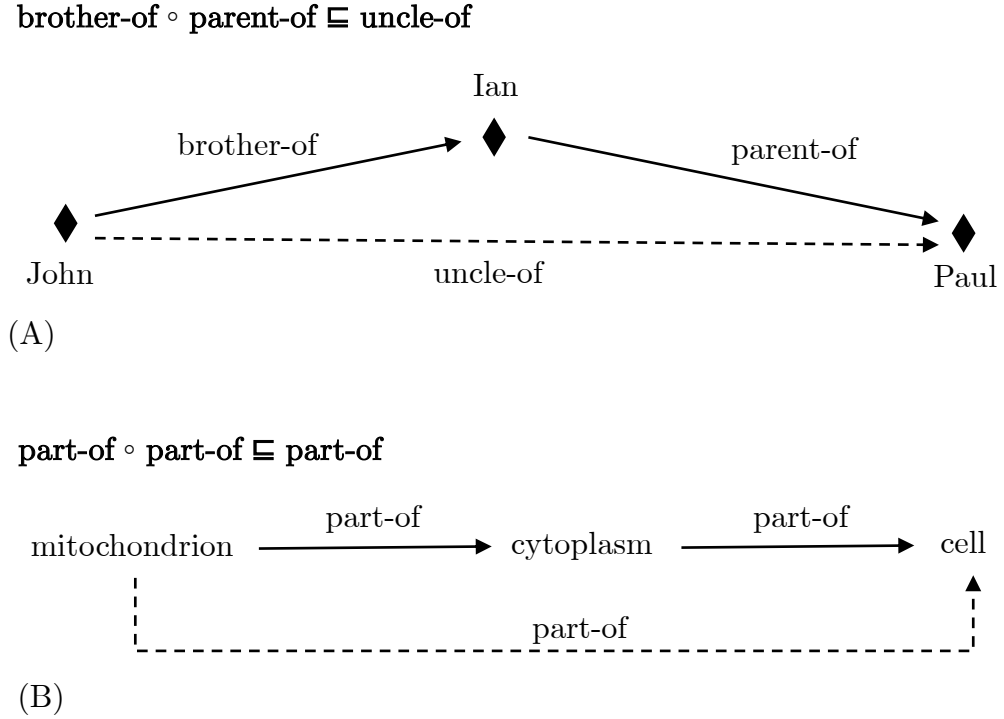


Figure 2.12: Description logics constructors: role composition. (A) Example of composed property, the *uncle* relationship. When two instances are linked using a *brother-of* property, then followed by a *parent-of* property, then a reasoner can create a property between the first and the last individual, as shown on figure between John and Paul. (B) Informal illustration of a transitive property using the GO specification (GO, 2014). When two terms are connected by a *part-of* relation and directly followed by another *part-of* property, then a *part-of* relation can be added between the first and last term. For instance here one can deduce that mitochondrion is part of cell, from the asserted facts. Note that this representation is not formal; OWL representation of OBO ontologies will be addressed later in this chapter.

2.3.4 Reasoning services

Combining basic entities, constructors and axioms gives a knowledge base or formal representation of a domain of interest whose content can be handled by a

reasoner. Considering the equation $x + 2 = 6$ as an analogy, axioms and entities helped to mathematically formulate the meaning of the symbols $=$, $+$ and 21; now a reasoner can automatically solve the equation and find the value of x. In the biomedical case, the problem faced is different; the main task of the reasoner will be to classify the knowledge base and to answer queries about the functioning of the molecular machine.

As briefly mentioned earlier, reasoners perform two types of operations: subsumption and consistency checking of the knowledge base. I will not discuss consistency checking, as it is only meaningful if a certain type of axiom is present (disjunction axiom - not presented in this document, yet part of the $\mathcal{EL}^{++}$ profile). The subsumption service is also called *classification*. In this context, it means assigning the right class to the correct place based on the meaning of the axioms. Classified classes form a *taxonomy*; for example the section 2.3.3.3 explains how the class *Father* can be asserted as equal to the intersection of the classes *Parent* and *Man*. From this axiom, the reasoner can deduce that *Father* is subsumed by Man (all fathers are man), therefore *Father* can be classified as subclass of *Man* and represented as such in a taxonomy (see Figure 2-13).

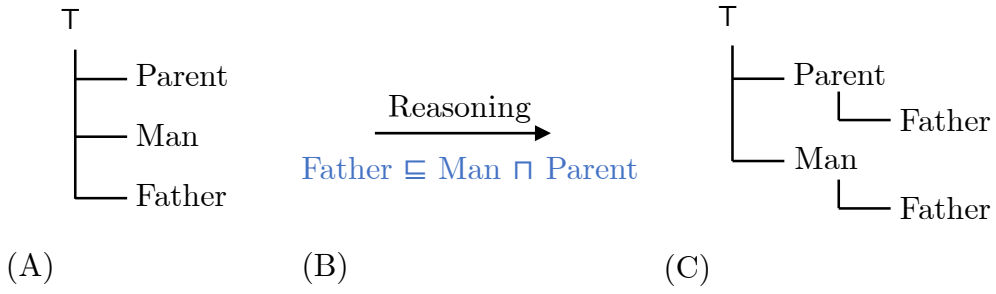


Figure 2.13: Description logics subsumption service example. (A) Taxonomy of classes not classified. (B) From the axiom present in the knowledge base *Father* $\sqsubseteq$ *Man* $\sqcap$ *Parent* (blue), the reasoning can deduce the taxonomy presented in (C). The class *Parent* and *Man* subsumes *Father*, which appears deeper in the hierarchy, evidence of a more expressive meaning. The symbol $\top$ represents the top class (*Thing* in OWL) and is always present above every other class of the knowledge base.

The subsumption service is also responsible for answering the queries formu-

lated over the knowledge base. A query is a standard class expression. The list of classes subsumed by the expression is the answer. DL queries are best orally formulated in the form of "What are the things that...". For example, "What are the things that are part of the cell?", expressed $\exists$ *part-of.Cell* in DLs. From the example of section 2.3.3.3, the reasoner would deduce nucleus (see Figure 2-11).

The complexity of reasoning services depends on the types of axioms present in the knowledge base. The ones I have discussed here are part of the $\mathcal{EL}^{++}$ profile, guaranteeing a tractable reasoning capable of handling large biomedical data input. Many more axiom types exists, featured by other DL families (Krötzsch, 2012). Deductions over the knowledge base are performed by the reasoner, and can be simplified as a classification where the taxonomy of classes is generated.

2.3.5 The Web Ontology Language 2 (OWL2)

DLs are a theoretical mathematical framework. It is possible to manually exploit the meaning of axioms in order to perform reasoning services, but the goal is to eventually use a computer to perform this task, for time and data size concerns. In this regard, DLs are specified for computer implementation by the Web Ontology Language (OWL2) (W3C, 2014c). OWL2 was primarily designed to help data interoperability over the World Wide Web and is tightly linked to the semantic web. Yet as OWL2 derives directly from DLs, it is possible to use the language to deal with DL-related problems, such as the study of the molecular black box machine. OWL2 terminology is slightly different from that of DLs, and the language provides some extra functionalities relevant to software implementation: concepts are called *classes*, and rather than having a human readable name such as *Father*, entities are identified with a Uniform Resource Identifier (URI), for example `http://www.example.org/Father`. This feature guarantees the provenance of the information, as domain names are unique in the World Wide Web (Berners-Lee et al., 2001). I will not discuss in details here the semantic web principles and design; the reader can simply consider that OWL2 enables the implementation of DLs in a computing setting.

One of the distinctive features of OWL2 is the *open world assumption*. This statement implies that nothing can be deduced from missing information. As an example: if the fact that drug A perturbs protein B is present and someone asks "Does drug A perturbs protein C?", the answer would be unknown. On the

contrary, in a closed world setting such as relational databases the answer would be "drug A does not perturb protein C". The open world assumption fits the requirement of biomedical knowledge well; it is fair to assume that our knowledge of the living world will never be complete, therefore deductions can only be made over explicit evidence.

2.4 Implementation with life-science information

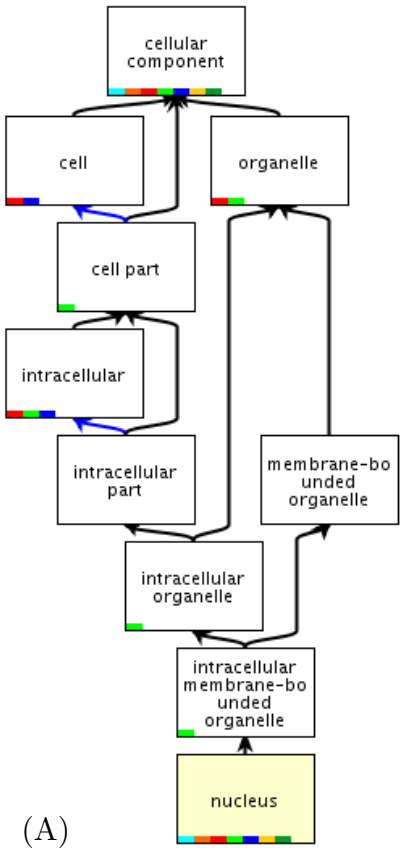
In summary, the theory presented previously states that DLs are a suitable theoretical framework for the descriptive study of biological organisms. The reader now might be wondering how an actual knowledge base is built using DLs. First of all, multiple implementations of the descriptive framework to study the black box machine can exist. The implementation of the theory varies depending on the questions asked and granularity required. Chapter 3 will present an example of implementation, made with a particular set of axioms, addressing the mode of action and drug repositioning. Other possible implementations could use other DLs features to characterise the black box machine to study other biomedical topics. However, in any case and in order to be successfully implemented in practice, this descriptive approach needs to extend the current solutions used to store biomedical data as much as possible. Reusing the information already available is beneficial as it decreases the amount of work to be done and allows for a non-disruptive transition from existing and adopted technologies. I will discuss here how the theory, namely the descriptive approach based on DLs, relates to current biomedical databases and ontologies - traditional keepers of the knowledge. Brain, a programmatic library dedicated to the OWL2 EL profile, will finally be introduced to show how scalable real-life applications can be built using DLs.

2.4.1 Integration with biomedical ontologies

2.4.1.1 Open Biomedical Ontologies (OBO)

In the early 2000s, with the advent of high-throughput DNA sequencing technologies, it became necessary to annotate genomes in a consistent fashion. The idea was to transfer the findings made in the sequences of one species to another organism. The Gene Ontology (GO) was first developed to address this prob-

lem (Ashburner et al., 2000) as a controlled vocabulary, representing molecular functions, biological processes and cellular locations. Following the successful adoption of the resource by the community, the Open Biomedical Ontologies (OBO) consortium was created, with the aim of creating a suite of orthogonal interoperable reference ontologies in the biomedical domain (OBO, 2014). OBO ontologies, usually contain synonyms of the described concepts as well as logical links between terms. It is encouraged to re-use the terms present in existing OBO ontologies, in order to create a net of biological concepts, sometimes called an integration layer. The OBO file format is traditionally used to serialise such ontologies (see Figure 2.14). This format provides a straightforward graph representation, similar to semantic nets, ancestors of DLs.



(A)

[Term]
id: GO:0005634
name: nucleus
namespace: cellular_component
def: "A membrane-bounded organelle of eukaryotic cells in which chromosomes are housed and replicated. In most cells, the nucleus contains all of the cell's chromosomes except the organellar chromosomes, and is the site of RNA synthesis and processing. In some species, or in specialized cell types, RNA metabolism or DNA replication may be absent."
is_a: GO:0043231 ! intracellular membrane-bounded organelle

(B)

Figure 2.14: Open Biomedical Ontologies format and representation. (A) The hierarchy above the term nucleus is presented. Black arrows entail an *is a* relation, blue ones a *part-of* relation. OBO ontologies are represented as directed acyclic graph. (B) Entry for the term nucleus, illustrating the format used to serialise OBO ontologies.

The simplicity of the OBO format certainly helped the adoption of the stan-

dard by the community, yet it presents a limited expressivity with loose semantics, in particular from a computational perspective. Following the recent rise of semantic web technologies, the OBO community is currently considering a shift from the original OBO format into OWL2. Automated conversion from one representation to another has been described in the literature (Tirmizi et al., 2011) (Hoehndorf et al., 2010). The main advantages in favour of adopting OWL2 are the possibility to reuse the numerous tools that have been developed as well as taking advantage of an extended expressivity to describe complex concepts.

The descriptive methodology presented in this chapter relates to the work done by the OBO community. First of all, as OBO semantics can be represented in OWL2, it is possible to directly reuse any OBO ontology to study the biological machine. Abstract concepts such as cellular functions or anatomical parts are extensively characterised inside some of the OBO ontologies and are therefore available to describe the parts of the cellular machinery. The main difference between OBO ontologies and descriptive knowledge bases comes from the lack of commitment towards interoperability and universality by the latter. Indeed, the reader will notice that I did not employ the word *ontology* to characterise the entity behind the study of the molecular machine. I purposefully avoided the use of the term in order to make clear that absolutely anything can be represented with DLs, as long as it serves somehow the study of a biomedical problem; the term *knowledge base* was used rather than *ontology* for this very reason. The main motivation behind the descriptive framework is to use computers and automated reasoning in an efficient, realistic and scalable fashion in order to derive biologically relevant information. Ontological commitment, in particular to top level concepts, entails the addition of complex axioms, which are hard to compute and most of the time outside of the OWL2 EL profile, such as universal quantifications for example (Krötzsch, 2012). As a result, it becomes practically and theoretically impossible to use any reasoning engine over such an input size, defeating the purpose of representing information using an expressive language as OWL in the first place.

2.4.1.2 Approximations and assumptions

The biomedical ontology community is very eclectic, with representatives from various disciplines such as mathematics, computer science, philosophy and bi-

ology. Despite enriching the dialogue and improving the overall quality of the work resulting from this complex social network of people, it can also sometimes lead to animated arguments, pitting the vision of one branch of the community against another (personal experiences). Ironically, the ontology community cannot even agree on the meaning of the word *ontology* (Schulz and Jansen, 2013). In order to still make interesting usage of the powerful framework of DLs, I came to consider a set of rules of thumb, identified throughout my work, which will be briefly summarised in this section.

OWL and DLs can only define things arbitrarily. Just as in statistics, which are very popular in biology, the tests and data representation considered always depend on the type of questions to be answered, balanced with their computational complexity. It has been extensively argued that DLs are only relevant for a subset of biomedical knowledge, in particular to handle statements that are so called *universally true* (Schulz and Jansen, 2013), meaning true all the time, no matter what. Approximations are strongly discouraged in this mindset. I personally came to disagree with this argument; heuristics and approximations are present everywhere in biology. For example, the typical biological axiom considered as universal truth is the subclass relationship between species (Schulz and Jansen, 2013) (Krötzsch, 2012).The class *Human* is a subclass of *Mammal*, meaning that every single human individual is also a mammal. A similar pattern can be applied across all other species. Any biologist will certainly agree that the boundary between two species is often fuzzy, as the main criteria used to distinguish them is based on the way organisms reproduce (Hanage, 2013). For some cases, such as mammals, it works rather intuitively well, but the approach is limited for some other species such as bacteria for instance (Hanage, 2013). This ambiguity is well known as the species problem and illustrates why it is impossible to simply clear cut two groups of living individuals in an ontological way while capturing evolution and biological meaningfulness. Another axiom widely considered as universal is the disjunction between the *Male* and *Female* classes, meaning that an individual cannot be a member of these two classes at the same time (W3C, 2014a). Such a statement, in addition to being socially offensive, also overlooks the complexity of the development of the hormonal system. Numerous existing individuals express sexual hormones in a deregulated fashion (1/1000 births) (Dreger, 1998), leading to conditions such as hermaphroditism or gonadal dysgenesis. Their sexual anatomy therefore spans in between the typically ex-

pected sexual traits. Such cases are particularly relevant to medicine, yet a naive ontological separation between sexes would classify these instances as inconsistent and not handle them properly.

As shown, modelling complex systems such as biological organisms almost necessarily requires some approximations. Universality rarely holds in nature; I therefore argue that it is acceptable to represent biological approximations in order to tackle relevant problems of interest. The main motivation behind this deductive framework is the attractive possibility to use it in combination with a computer; this characteristic makes DLs unique and the focus should be put there. Biomedical axioms will most likely always be approximations, because of the underlying physical complexity of biological systems. By using the less restrictive term *knowledge base* rather than *ontology* I also wanted to make this distinction: biomedically relevant formal deductions can be made using DLs, and biological approximations can be represented. The rules of thumb to guide the representation I have adopted are the following:

- Defining a series of competency questions. Such questions are the tests a reasoner should be able to deduce from a formal knowledge base. All the modelling resolves later to answer correctly these questions of interest. Approximations can be made, as long the assumptions are understood and interpreted correctly.

- Staying in the $\mathcal{EL}^{++}$ (or OWL2 EL) profile, which guarantees scalability (Hoehndorf et al., 2011a). Nowadays, the input sizes of interesting biomedical problems are too large for more expressive families such as OWL2 Full or OWL2 DL. Examples from the past show how a too expressive modelling fails to scale and give the promised answers in terms of reasoning (Vempati et al., 2012) (Golbreich et al., 2006) (Mungall et al., 2010) (Mungall et al., 2011) (Villanueva-Rosales and Dumontier, 2008).

- Trying to compromise between the constructs available inside $\mathcal{EL}^{++}$ and biomedical approximations, in order to answer the competency questions.

- Not aiming for universality. It is of good practice to reuse and scale over the work done by other people, in the quest of interoperability (reusing terms, identifiers, patterns, etc.). However, the race for universality is much more challenging, as discussed before. I always advise designing knowledge bases

to first answer the competency questions, and secondly, depending on time and energy, to maximise interoperability and aim for universality.

- It is to be assumed that deductions made over a biomedical knowledge base still require human interpretation. OWL and DLs are sometimes marketed (including by myself) as "knowledge discovery" tools; in practice the true discovery of new knowledge is made by a biomedical researcher, guided by the content of the knowledge base. A representative illustration of this argument is a figure featured on the original article introducing the GO (Ashburner et al., 2000), where the authors erroneously draw *is a* arrows linking terms in the opposite direction. Despite being semantically erroneous, the logic can be easily interpreted by any biologists, hence the success of the resource.

- Finally, I quote Hendler (Hendler, 2014): *a little semantics goes a long way*. This sentence is the motto of the semantic web movement and is widely accepted in the biomedical community too. By focusing on a few axioms, in particular on relational ones, a great and concise expressivity can be reached, as it will be shown with the definition of the mode of action presented in the coming chapter.

2.4.2 Integration with databases

As of 2014, a large amount of biomedical information is stored inside relational databases (Brooksbank et al., 2014). Some of this content is publicly available over the Internet, distributed by large organisations such as the European Bioinformatics Institute (EBI) or the National Center for Biotechnology Information (NCBI). Each database usually focuses on one theme in particular: for instance, ChEMBL (Gaulton et al., 2012) provides millions of records about small molecules and their bioactivity against protein targets. These protein targets are themselves referenced inside Uniprot (Consortium et al., 2013), a resource indexing the known information related to gene products. Records described in one place are moreover hyper-linked or crossed-referenced to records present in another resource, with the help of identifiers.

Biomedical databases can be seen as catalogues indexing the parts of the molecular machinery, yet not providing any logical information on how these en-

tities connect. Interestingly, DLs provide the means to capture this logical layer, thanks to the expressive power of roles (object properties in OWL2). DLs can perfectly integrate with the information currently provided by biological databases, in order to enhance the semantics of the data. Consider the following scenario as an example: it is known from experimental evidence that a protein X is involved in blood coagulation, and a researcher would like to record this biological piece of knowledge. As of 2013, this statement is partially captured using the information of three repositories: Uniprot, the GO and the GO Annotations (GOA). The database Uniprot provides the identifier for the protein X, the GO gives an identifier for the term *blood coagulation*, and finally inside GOA one would find an association (pairing) between protein X and *blood coagulation*. From the DLs perspective, the same statement can be formally captured by an axiom: *protein X subClassOf involved-in some Blood Coagulation*. Here the logic-less annotation or association between the protein and the biological process has been formalised into something more meaningful and expressive. The role *involved in* can indeed be combined with other roles in order to reflect the logic behind the biological system. This example shows how DLs can reuse the information already present and leverage the connectivity between classes (Jupp et al., 2012). I will present in the next chapter how these relations can be combined and structured to define the concept of mode of action and automatically classify drugs.

Converting the information from a relational database into OWL implies a change in the representation. Traditionally, database entries are considered as instance records, belonging to a table or schema. This fact implies that a record, such as protein X as available in Uniprot would correspond to an instance or individual in DLs (section 2.3.3.1). However, in practice, billions of copies of this canonical protein X exist. Therefore when biological entities or concepts are modelled in OWL, instances often become classes. This representation is closer to the reality and should simplify the modelling and connection with other concepts, like molecular functions. However, it is perfectly acceptable to represent proteins as instances too in my opinion, depending on the type of question being asked over the knowledge base. For instance, proteins could be individuals within a protein or domain family, as indexed by InterPro, database dealing with protein domains (Zdobnov and Apweiler, 2001).

2.4.3 Brain library - implementing programmatic solutions

Successfully implementing a knowledge base requires a compromise between scalability and expressivity. I argued in favour of the $\mathcal{EL}^{++}$ profile (equivalent to OWL2 EL), providing an adequate expressivity for biomedical sciences and enabling tractable reasoning. The discussion so far was mostly focused on the theoretical aspect, yet in order to be used in a programmatic way, a library or programmatic framework is needed. At the time of writing, two main free and open-source solutions exists to work with the OWL2 EL profile: Protege (Knublauch et al., 2005) and the the OWL-API (Horridge and Bechhofer, 2011). Protege is a popular graphical user interface, useful to develop toy examples and define the core axioms of a knowledge base. However, it is not very suitable for large knowledge bases, where potentially thousands of classes need to be handled. A programmatic alternative is a Java-based library called OWL-API. The framework implements the standard specification for OWL2 in deep granularity, yet it becomes quickly cumbersome to work with it in order to perform analyses and run biomedical queries. For these reasons, I developed Brain (Croset et al., 2013a), a Java library bridging the gap between these two solutions. This work is freely available online (`https://github.com/loopasam/Brain`) and open source. The library features a simplified interaction with DLs and OWL2 axioms using the Manchester syntax (Horridge et al., 2006). Figure 2-15 shows an example of a program written with the library.

Brain uses the ELK reasoner (Kazakov et al., 2013) to perform reasoning services over the knowledge base. ELK is dedicated to the EL profile, fast (Gonçalves et al., 2013) and can perform reasoning tasks in parallel; these characteristics make it an ideal candidate to handle biomedical knowledge. The Brain library mostly focuses on querying the underlying knowledge base; in this regards, OWL queries can be formulated as string expressions, which will then get automatically converted into Java objects and processed by the reasoner. Web applications can be safely built over the framework, as special care was put on thread management and coordination. Brain provides as well a series of convenience methods, useful to address biomedical questions; indeed, biological inference often derives from similarity metrics, such as sequence comparison (Stevens et al., 2007); the taxonomic structure of a knowledge base can also be used to derive a closeness

```
//Creation of the Brain object instance
Brain brain = new Brain();

//Add an OWL class to the knowledge base:
brain.addClass("Nucleus");
brain.addClass("Cell");

//Add an OWL object property:
brain.addObjectProperty("part-of");

//Declare the axiom:
brain.subClassOf("Nucleus", "part-of some Cell");

//Integrate the content of an external knowledge base:
brain.learn("http://example.org/bar.owl");

//Query the knowledge base:
List<String> subclasses =
                    brain.getSubClasses("part-of some Cell", false);

//Free the resources used by the reasoner:
brain.sleep();

//Save the knowledge base:
brain.save("/path/to/ontology.owl");
```

Figure 2.15: Example of Java program written using the Brain library. Each command is preceded by a comment explaining the functionality.

index, so called semantic similarity, reflecting how close two entities are in the classification. The convenience methods provided by Brain calculate the Jaccard index over the set of superclasses. An illustration of the methodology is depicted in Figure 2-16. This type of analysis is out of the scope of DLs, yet particularly important for generating and exploring drug repurposing hypotheses, and will be further discussed in the coming chapters. The library offers the possibility to export graphs of the knowledge base, as exemplified by Figure 2-17. This functionality comes in handy to build web applications and to present content to users. Finally the reader might be interested by Tawny OWL (Lord, 2013), a similar application released approximately at the same time as Brain and with similar goals, but written in Closure.

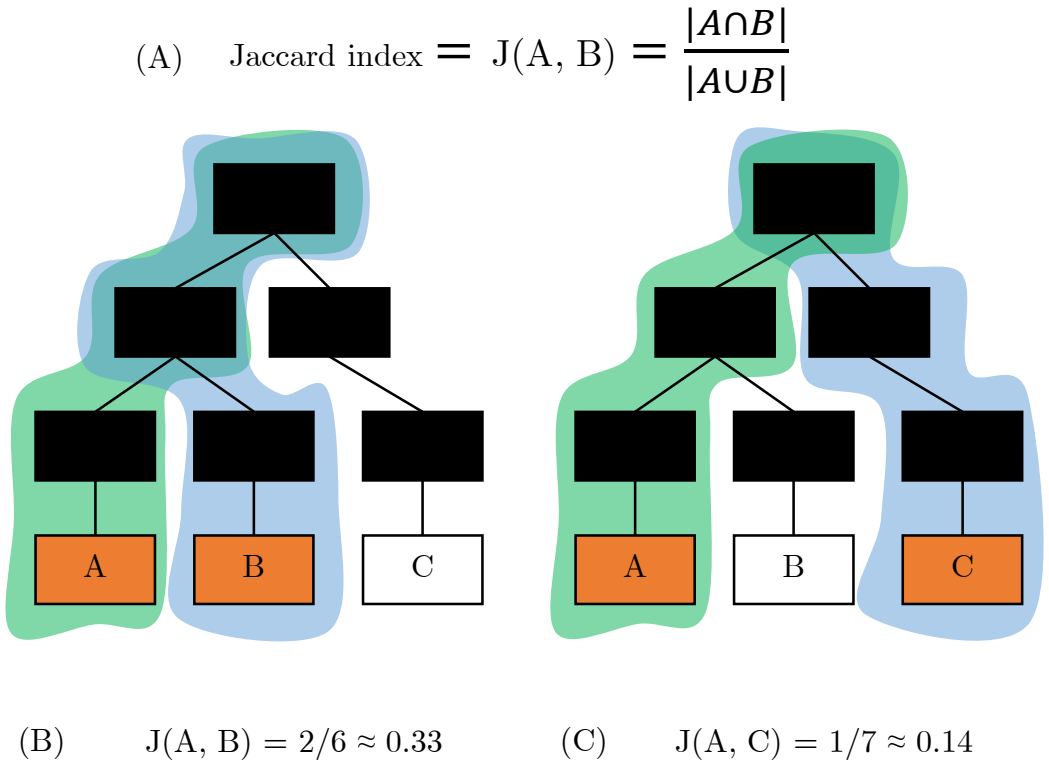


Figure 2.16: Jaccard coefficient implementation in the Brain library alongside examples. (A) Definition of the coefficient: The similarity between two entities is defined as the ratio of the categories in common divided by the total number of categories. (B) and (C) Examples of computation of the index over two pairs (letters and in orange). The taxonomy is in black, super-classes shown with the blue or green area. The index is higher between the pairs A and B (0.33) than between the pair A and C (0.14), representing the fact that A and B are closer in the taxonomy and have a more similar meaning. The coefficient will be used greatly later to compare the function of drugs.

2.5 Summary

Living organisms are complex, yet understanding their internal machinery is required in order to develop new therapies and find treatments for diseases. Despite being eventually only made of chemical interactions, biological systems can be best analysed at a higher abstraction level. I presented how the study of life and resulting knowledge can be formalised using DLs. The main advantage of this framework is the possibility of defining abstract concepts (e.g, processes or phenotypes) as well as real entities (e.g, proteins or metabolites) in order to query or

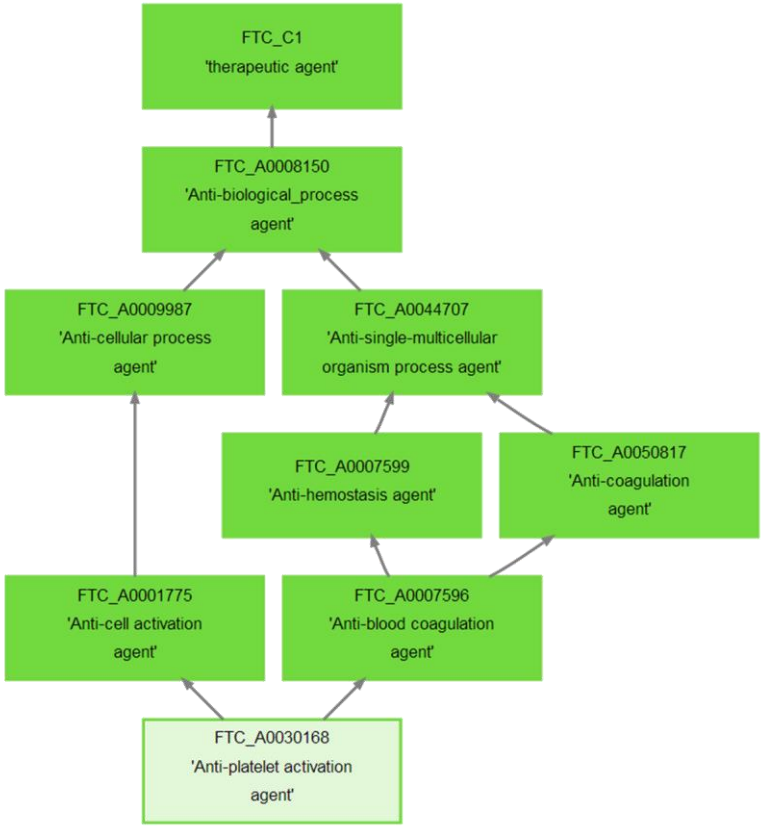


Figure 2.17: Example of graph generated from the Brain library (content not pertinent): from an input concept, it is possible to export the whole ascendant taxonomy as scalable vector graphics (SVG), particularly useful to display information on web browsers.

derive information using a computer. The logical links between modules of the cellular machinery can be modelled, and, computing their entailments resolves to a classification problem. DLs are less precise than molecular modelling, yet it is possible to derive formal solutions while considering the cellular system as a whole.

This approach integrates nicely the current information, available in resources such as databases and ontologies, in order to derive practical implementations. The analogy cell-machine will serve as the theoretical basis to formally define the concept of mode of action and address drug repurposing questions, as I will show in the next chapter and as summarised in Figure 2-18. Scalability was a core concern, in order to implement robust application. In this regard, I discussed the $\mathcal{EL}^{++}$ profile, designed and inspired by the axiom types found in biomed-

ical ontologies like SNOMED and guaranteeing the implementation of realistic solutions.

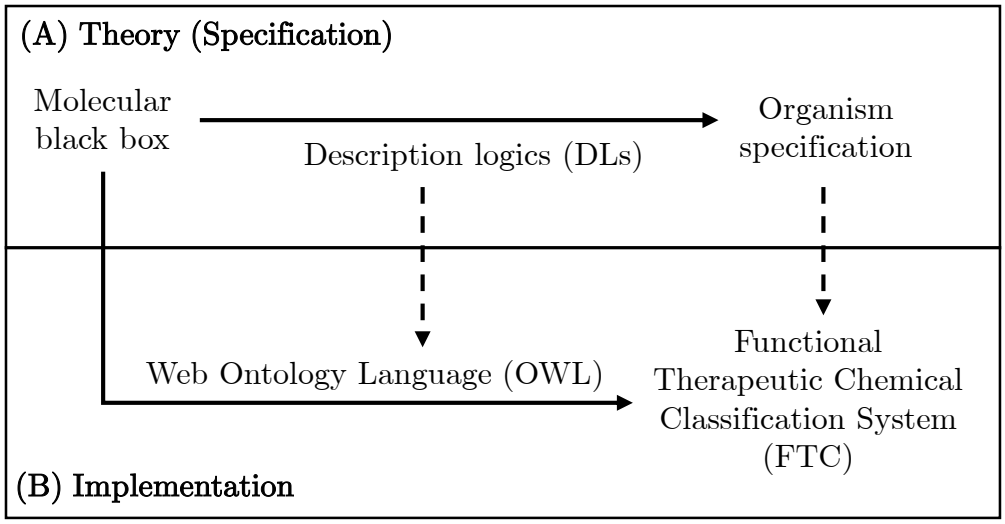


Figure 2.18: Summary of the molecular black box theory. (A) To address drug discovery, organisms can be compared to black box machines. Description logics are a useful framework to unveil the specification of the black box machine. (B) The theory can be implemented (dashed arrow) for computers using the Web Ontology Language (OWL). Chapter 3 will present an example of implementation of the theory with the Functional Therapeutic Chemical Classification System (FTC), dedicated to handle the concept of mode of action and suited to repurpose drugs.

chapter 3

The Functional Therapeutic Chemical Classification System (Implementation)

Key points

- New mode and mechanism of action (MoA) concepts can be formally created using description logics (DLs) and following the principles introduced in Chapter 2.
- 20,000 new MoA concepts are created and present in a novel resource called the Functional Therapeutic Chemical Classification System (FTC). The resource has a taxonomic structure and describes the biological roles of drugs (e.g. *anti-blood coagulation agent*).
- Over a thousand of approved drugs are classified inside the FTC categories by integrating the content of DrugBank, UniProt and Gene Ontology Annotations (GOA) and with the help of an OWL reasoner. The classification process is fast and complete, as the axioms present in the FTC follow the $\mathcal{EL}^{++}$ profile.
- The biomedical information present in the FTC is evaluated against the content of the Anatomical Therapeutic Chemical Classification System (ATC), manually curated resource describing the indication, function and therapeutic areas of approved drugs. Briefly, the drugs classified in the FTC

are covering 89% of the content of the ATC (recall). Given a therapeutic category, the FTC contains more drugs than the ATC, reflecting the drug's polypharmacology (precision of 50%).

- The content of the FTC will be used to derive systematic drug repositioning hypotheses as presented in Chapter 4.

Author's comment

The content and structure of this chapter were directly extracted from the published article (Croset et al., 2013b) describing the FTC and some of the analyses performed over it. I have edited the text in order to provide more details when needed and relevant, and to hopefully keep the coherence with the rest of the document. The argumentation has a classical structure: The methodology behind the creation of the resource is first presented, followed by an evaluation section. The results obtained are finally contrasted and discussed.

3.1 Introduction

Chapter 1 introduced my thesis, namely formally representing drug's mechanisms and modes of action in order to discover new indications. In Chapter 2 was presented a theoretical perspective on description logics (DLs) and their relation to the study of life. This chapter implements the theory and describes the generation and evaluation of MoA categories.

As stated in Chapter 1, drug repurposing is the use of known active compounds for new therapeutic indications (Sanseau and Koehler, 2011). When administered in a living organism, a compound can indeed play various roles and affect different biological processes; accurately identifying these different functions helps to predict the potential side-effects a drug can have and can also lead to interesting repositioning opportunities (Medina-Franco et al., 2013). For instance, sildenafil was initially developed to relieve angina pectoris symptoms and has been repurposed towards erectile dysfunction during the clinical trials when a new function of the target enzyme was discovered (see Chapter 1 section 1.2.1).

Approved compounds are attractive because they have been extensively studied and have by definition already successfully passed clinical trials, where most

drugs fail because of safety or efficacy issues. There is an increasing number of approaches to predict repurposing opportunities using computational methods (see Chapter 1). Most methods operate on the profiles of physicochemical descriptors derived from molecular structures (Haupt and Schroeder, 2011). Other methods characterise the drugs on more abstract levels, such as the gene expression signature (Iorio et al., 2010) or via the reported side-effects (Campillos et al., 2008). These approaches have in common to look for similarities within existing drugs and forward similar compounds as repositioning hypotheses.

A feature of particular interest to describe drugs is the MoA. According to Wikipedia, the MoA describes *a functional or anatomical change, at the cellular level, resulting from the exposure of a living organism to a substance.* For instance terms such as *transcriptional regulation agent* or *anticoagulant* define MoAs and characterise the roles of a certain type of drug. The MoA abstracts over the relations between molecular functions, protein targets and drug activities; it is the central concept linking a chemical structure to a set of biological activities (see Chapter 1). Intuitively, the indication of a drug logically depends on its MoA (see Figure 3.1).

Despite its widespread use in drug discovery, the MoA has not been used yet as a descriptor for repositioning analyses. One reason for this might be the challenge of formally defining the concept. Indeed, MoAs are terms or categories, it is not possible to represent them straightforwardly with values and numbers like one can do for a 3D molecular structure or for a gene expression profile. Nonetheless, the meaning of a concept can be formalised with controlled vocabularies and ontologies (Gruber et al., 2009); such frameworks help to formalise the semantics of symbols and strings of characters with explicit axioms (see Chapter 2).

In an ontology or knowledge base, concepts (interchangeable with *category*, *term* and *class* in this document) are organised and linked following the logical type of relation they have among them. In the Gene Ontology (GO) for example, biological processes and molecular functions terms are manually curated and their meaning specified by the relation types linking two GO terms. MoA definitions are present in other classifications such as the Medical Subject Headings (Nelson et al., 2004) or the Chemical Entities of Biological Interest (Hastings et al., 2013). The Anatomical Therapeutic Chemical Classification System (ATC) (Organization et al., 2006) also describes to some extent the action of drugs at the anatomical level. All these resources are valuable for the community as a source

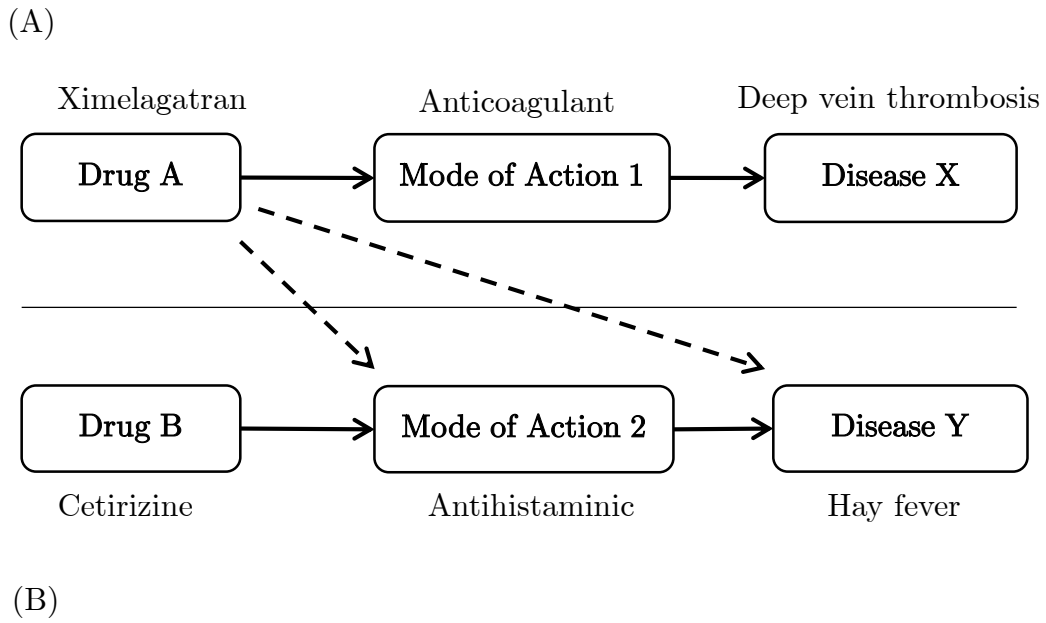


Figure 3.1: Conceptual relationship between a drug, its mode of action (MoA) and disease indication. Because a drug exhibits a certain MoA it is therefore indicated for a diseases, as showed by examples (A) and (B). If a new MoA was discovered for a drug (e.g. ximelagatran with antihistaminic MoA), the compound could be re-indicated accordingly (in this case for hay fever illustrated by the dashed arrows). In order for such a deduction to be made, MoA categories first have to be represented, and secondly drugs have to be assigned to these categories.

of carefully manually curated information. Moreover, the categories described in these classification systems are sometimes used to annotate drugs: for instance the compound sildenafil has been manually annotated as *vasodilator agent* (CHEBI:35620 or MeSH:D27.505.954.411.918).

The classifications mentioned previously are not specially designed for drug repositioning; they purposefully report only the well-known and major MoAs of chemical compounds. The pharmacological spectrum of each drug is not necessarily well covered, yet it would be the best way to predict new indications. In my context, an ideal knowledge base would feature the known MoAs of a drug as well as some predicted ones to be tested in experiments. The MoA categories should derive and scale over primary molecular evidence exposed in biomedical databases, in an automated way as motivated in Chapter 2.

To address the lack of systematic MoA annotations, I have implemented the Functional Therapeutic Chemical Classification System (FTC), presented here in

this chapter. The FTC is automatically built by leveraging the content of various biomedical databases using DLs and automated reasoning. Over 20,000 new MoA categories are defined in the resource and further populated with approved drugs using the Web Ontology Language (OWL) in combination with a reasoner. The population step takes in account the type of pharmacological action, the molecular targets of the drugs and their involvement into multiple biological processes.

Drugs can exhibit several MoAs, and the same MoA can be reached through different mechanisms. Most of the drugs are present in multiple FTC categories, reflecting the various roles a compound can play inside a biological system which can serve as starting point for drug repositioning. The resource was evaluated against the ATC, traditional classification scheme introduced before. I present as well some preliminary analyses over the data, by looking at the relation between the MoA and the indication of a compound using semantic similarity. Finer analyses and repositioning use-cases such as Alzheimer's disease and hypertension will be investigated in Chapter 4.

3.2 Method and definitions

This section describes the building mechanism behind the FTC. The full list of axioms composing the knowledge base are listed at the end of this section (section 3.2.9). The FTC is one possible implementation of the theory described in Chapter 2, dedicated to handle MoAs.

Summarised, the creation of the FTC follows these steps: first, a list of categories describing the mode and mechanism of action of drugs is defined. Then in a second step the newly created categories are automatically populated with approved compounds. Finally, the FTC is evaluated and repositioning hypotheses can be generated (presented in Chapter 4).

3.2.1 Source code

The code behind the creation of the resource is entirely open and available at `https://github.com/loopasam/ftc`. The web application built on the top of the FTC can be found at `https://www.ebi.ac.uk/chembl/ftc` and the documentation can be accessed at `https://github.com/loopasam/ftc/wiki`. The reader should be familiar with DLs and the Web Ontology Language (OWL) to

fully understand the construction of the knowledge base. An introduction to DLs from the perspective of the biomedical scientist is available on the wiki at `https://github.com/loopasam/ftc/wiki/Description-Logics` and in Chapter 2 section 2.3. The FTC implementation relies mostly on Brain (Croset et al., 2013a) and the web application builds on the top of the Play! framework. Classification tasks use the ELK reasoner (Kazakov et al., 2013). The computer hosting the web application has 8 GB of memory with 4 processors, this architecture allows fast parallel reasoning, thanks to ELK's design. More functionalities will be added to the web application following user requirements (*lean implementation*).

3.2.2 Categories creation

The mode of action categories present in the FTC are defined based on the terms coming from the Gene Ontology (GO). Both the molecular function and biological process branches are used for this purpose, yet handled slightly differently as described below.

3.2.2.1 Mode of Action categories

All the biological processes featured in the GO are looked-up one by one. All the time a process is linked to another process (X) via a *positive* or *negative regulation* link, two FTC classes are created: *Anti-X agent* and *Pro-X agent*. For instance the GO term *positive regulation of blood coagulation* is linked to the term *blood coagulation* via a *positively regulates* relation, therefore two FTC categories *Anti-blood coagulation agent* and *Pro-blood coagulation agent* are created. The identifiers of the new FTC classes are also derived from the GO term used to create the class pattern. The GO numeric identifier is re-used and the letter A or P is appended before to emphasize the *anti* or *pro* pattern. From the example presented previously, the FTC class *Anti-blood coagulation* has FTC_A0007596 as identifier, because the GO term *blood coagulation* is referenced as GO:0007596. Following the same logic, FTC_P0007596 is the identifier of the class *Pro-blood coagulation agent*. The design choice for identifiers and labels allows the FTC to fully rely on the high quality work provided by the GO curation team and scale over it.

3.2.2.2 Mechanism of Action categories

The mechanism of actions related to molecular functions are created in the following manner: all the time a molecular function (Y) is encountered then two FTC categories are created, as for the processes: *Anti-Y agent* and *Pro-Y agent*. The identifiers are assigned the same way as described before. For instance, out of the GO term *androgen receptor activity* (GO:0004882) two FTC classes are derived: *Pro-androgen receptor activity agent* (FTC_P0004882) and *Anti-androgen receptor activity agent* (FTC_A0004882).

3.2.3 Equivalent definitions

FTC classes are generated as presented in the previous section. Up to this point, these categories are only tokens with a human readable label as well as an identifier. The next step is going to assign equivalent definitions to each FTC class. An OWL reasoner can understand such definitions and will automatically classify the knowledge base accordingly, following standard DLs reasoning services (see Chapter 2 section 2.3.4). Drugs will then be assigned to FTC categories and the taxonomic structure arises after this reasoning step. Equivalent definitions are written as OWL class expressions using the entities of the knowledge base (summarised at `https://github.com/loopasam/ftc/wiki/Knowledge-Base` and in section 3.2.9). There are two types of equivalences: the first one captures perturbation of regulatory biological processes (so called *regulatory patterns*) and the second one handles the perturbed functions (*functional patterns*).

3.2.3.1 Regulatory pattern

Some of the FTC categories are created from the biological processes present in the GO (cf section 3.2.2); these categories have two arbitrary equivalent definitions, representing the two possible ways a compound might impact the biological process. Anti-biological process agent FTC categories contain the drugs that negatively perturb a target involved in the positive regulation of the biological process. The *anti* categories also feature the compounds that positively perturb a negative regulator of the same process. The *pro* categories are equivalent to the opposite pattern. Figure 3.2 and 3.3 illustrates the equivalent definitions for the FTC classes *Anti-blood coagulation agent* and *Anti-blood coagulation agent*.

Anti-blood coagulation agent =

```
Drug and negatively-perturbs some (
  Protein and involved-in some (
    Biological-Process and positively-regulates some blood-coagulation
  )
)
```

```
Drug and positively-perturbs some (
 Protein and involved-in some (
    Biological-Process and negatively-regulates some blood-coagulation
  )
)
```

Figure 3.2: Regulatory pattern. Equivalent definitions for the concept *Anti-blood coagulation agent*. The concept is asserted as equivalent to either of the boxed expressions. A reasoner can understand such definition and classify drugs accordingly.

Pro-blood coagulation agent =

```
Drug and positively-perturbs some (
  Protein and involved-in some (
    Biological-Process and positively-regulates some blood-coagulation
  )
)
```

```
Drug and negatively-perturbs some (
 Protein and involved-in some (
    Biological-Process and negatively-regulates some blood-coagulation
  )
)
```

Figure 3.3: Example of regulatory pattern. Equivalent definitions for the concept *Pro-blood coagulation agent*. The concept is asserted as equivalent to either of the boxed expressions. A reasoner can understand such definition and classify drugs accordingly.

Figures 3.4 and 3.5 present the biological motivation behind the regulatory patterns: it should be easier to adjust the dosage for the compounds classified as such.

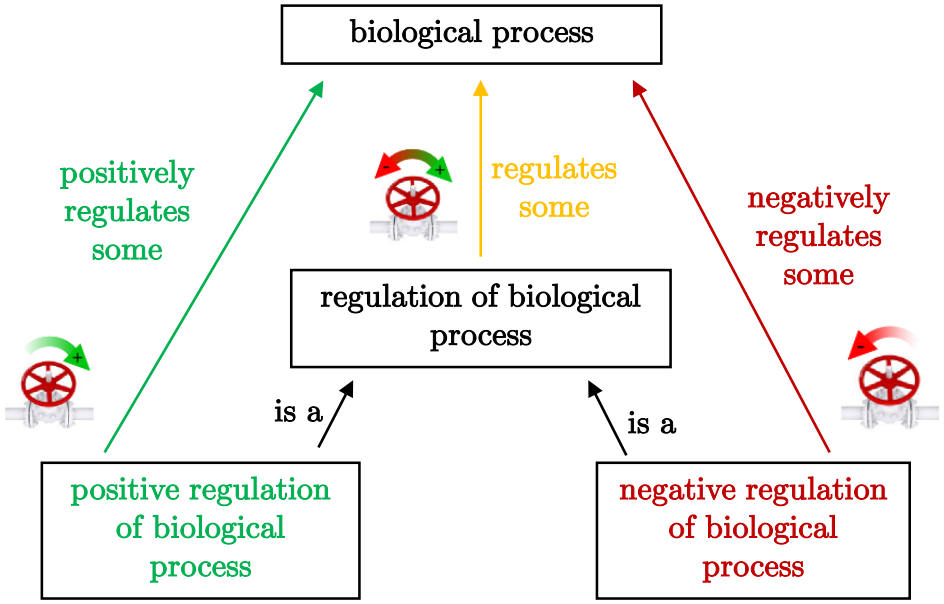


Figure 3.4: Biological processes of therapeutic interest are perturbed via regulators of the given process; this strategy allows to modulate and tune the effect, rather than blocking it totally. A regulatory process can be seen as a valve controlling the amplitude or frequency of another process, as defined in the GO. This characteristic is of interest for drug discovery, it means that the strength of the pharmacological effect is more likely adaptable with the dosage and drug's concentration.

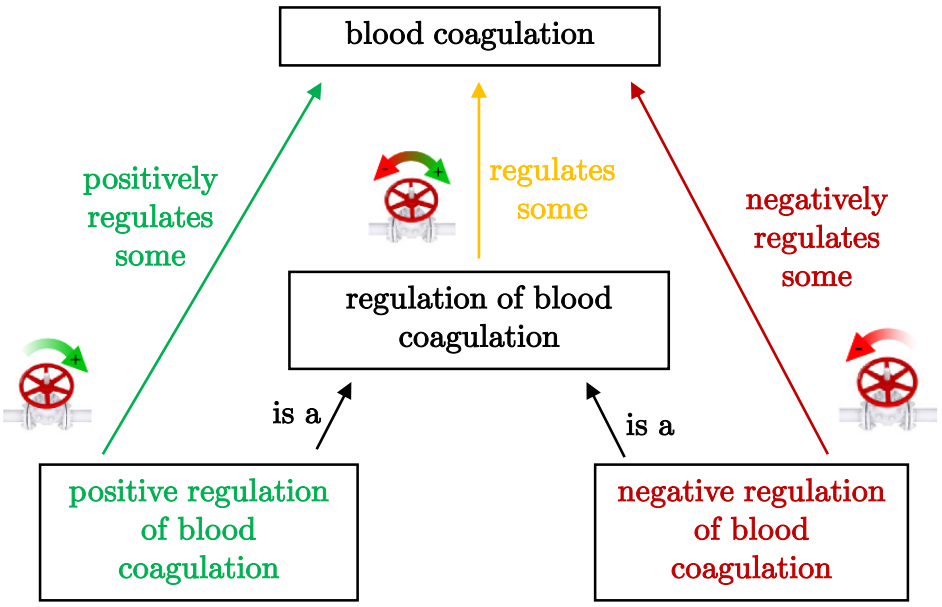


Figure 3.5: Example of regulation of the blood coagulation process, as defined in the Gene Ontology. Perturbing the coagulation via a regulator allows to more finely control the therapeutic outcome. See Figure 3.4 for theoretical illustration.

3.2.3.2 Functional pattern

The FTC categories generated from the GO molecular functions (cf section 3.2.2) are also equivalent to a logical definition. *Anti* FTC categories dealing with molecular activities are asserted as equals to the drugs that negatively perturb the function. *Pro* categories are equivalent to the drugs that positively perturb the function of interest. A summary of the functional pattern definitions is available on the online wiki at `https://github.com/loopasam/ftc/wiki/Mode-of-Action` and on Figure 3.6.

Anti-molecular function agent =

```
Drug and negatively-perturbs some (
  Protein and has-function some (
   molecular-function
  )
)
```

Pro-molecular function agent =

```
Drug and positively-perturbs some (
  Protein and has-function some (
   molecular-function
  )
)
```

Figure 3.6: Example of functional patterns; equivalent definitions for the concepts *pro* and *anti-molecular function agent*. These concepts is asserted as equivalent to either of the boxed expressions. A reasoner can understand such definition and classify drugs accordingly.

3.2.4 Data integration

At this stage, the knowledge base contains the created FTC classes associated with their logical definitions, as well as the GO and the core FTC entities. The knowledge base is then further populated with some information coming from various public databases. Only manually curated information extracted from peer-reviewed literature with experimental evidence is considered.

3.2.4.1 Drugbank

The DrugBank database is a unique bioinformatics and cheminformatics resource that combines detailed drug data with comprehensive drug target information (Knox et al., 2011). The approved drugs (small molecules and biotherapeutics) acting on proteins are extracted from the database and imported in the FTC knowledge base. In order to be selected, a compound must firstly be approved and secondly have a pharmacological action on at least one human protein target present in Uniprot (Consortium et al., 2013). The protein targets all have at

least one manually asserted GO annotation (Dimmer et al., 2012) for a biological process or a molecular function. DrugBank links compounds to targets via *actions*. The DrugBank actions are somehow structured and consistent: concepts such as *inhibitor* or *agonist* are reused throughout the database for example, yet they are not strictly formalised as a controlled vocabulary. These actions are manually standardised to the core properties of the FTC according to their biochemical meaning: for instance the action *antagonist* is mapped to the FTC *negatively-perturbs* property. The full list of mappings is available on Table 3.1.

DrugBank pharmacological action	mapped FTC property
agonist	'positively-perturbs'
potentiator	'positively-perturbs'
inducer	'positively-perturbs'
partial agonist	'positively-perturbs'
activator	'positively-perturbs'
ligand	'positively-perturbs'
stimulator	'positively-perturbs'
unknown	'perturbs'
cofactor	'perturbs'
other/unknown	'perturbs'
binder	'perturbs'
other	'perturbs'
allosteric modulator	'perturbs'
multitarget	'perturbs'
modulator	'perturbs'
partial antagonist	'perturbs'
Binder	'perturbs'
reducer	'negatively-perturbs'
inhibitor	'negatively-perturbs'
antagonist	'negatively-perturbs'
negative modulator	'negatively-perturbs'
cross-linking/alkylation	'negatively-perturbs'
intercalation	'negatively-perturbs'
adduct	'negatively-perturbs'
chelator	'negatively-perturbs'
antibody	'negatively-perturbs'
incorporation into and destabilization	'negatively-perturbs'
cleavage	'negatively-perturbs'
inverse agonist	'negatively-perturbs'
suppressor	'negatively-perturbs'
inhibitory allosteric modulator	'negatively-perturbs'
inhibitor, competitive	'negatively-perturbs'

Table 3.1: Mapping of DrugBank vocabulary to the FTC object properties.

Compounds coming from DrugBank are represented as OWL classes and asserted as subclasses of the class *DrugBank compound* (FTC_C2). Protein targets are described as OWL classes too and subclasses of the core class *Protein*. Each DrugBank compound is then connected to its target via the following axiom pattern: *[drug] subClassOf perturbs some [protein]*. E.g. *Ximelagatran subClassOf negatively-perturbs some Prothrombin*.

3.2.4.2 Gene Ontology Annotations (GOA)

The GO annotation program aims to provide high-quality GO annotations to proteins in UniProt (Dimmer et al., 2012). In the context of the FTC, such annotations are used to create axioms linking protein targets to molecular functions and biological processes. Each protein annotated with a function creates an axiom such as *[protein] subClassOf has-function some [molecular function]*. Each protein annotated to a biological process creates an axiom such as *[protein] subClassOf involved-in some [biological process]*. E.g. *Prothrombin subClassOf involved-in some positive regulation of blood coagulation*. Each protein can be involved in multiple processes and capable of performing multiple functions; some of the polypharmacology is captured at this level.

3.2.5 Knowledge base classification

The knowledge base is fully built at this step and contains core classes, MoA categories alongside the actions of approved DrugBank compounds on protein targets in Uniprot. The proteins are linked to their molecular functions and involvement in biological processes via the GO annotations. The logical specifications of the FTC are there to glue the different data together and to explicitly express the logical links between resources. The FTC knowledge base follows an OWL2 EL profile (implementation of $\mathcal{EL}^{++}$) (Motik et al., 2009), which enable the use of fast and parallelised reasoners such as ELK. During the classification process, the reasoner checks whether the MoA equivalent definitions are satisfied or not and assigns drugs inside the corresponding FTC categories. The tree structure of FTC appears also at this step from the logical definitions.

3.2.6 Evaluation methodology

As the classification of therapeutic agents is done in an automated way, it is important to evaluate the results generated against a known resource which will be considered as gold standard. The assessment of the FTC is done against another similar classification, the Anatomical Therapeutic Chemical Classification System (ATC) (Organization et al., 2006). The ATC *has been developed to serve as a tool for drug utilisation research in order to improve quality of drug use* (Organization et al., 2006). In this resource, the information is manually curated, and drugs are assigned to categories based of their legally approved indications. Figure 3.7 provides a summary of the classification as well as a reference explaining the different levels and their meaning.

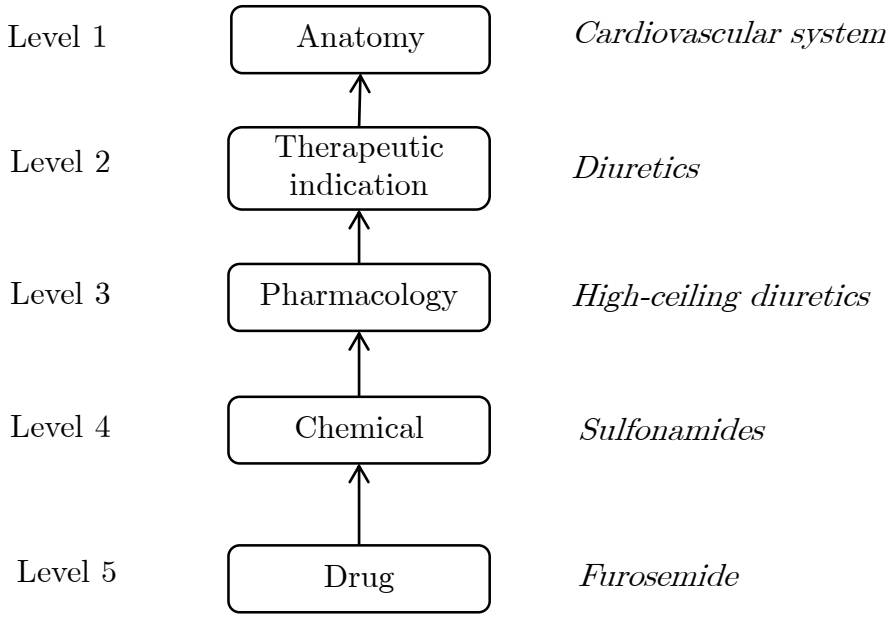


Figure 3.7: The structure of the Anatomical Therapeutic Chemical Classification System (ATC). The classification is composed of 5 levels. The first one describes the main anatomical group, the second one reflects the indication or therapeutic area of the drug. Level 3 handles the pharmacological action, level 4 describes the chemical structure, and finally level 5 contains drug's names. Examples are provided on the right column (italics) for the drug furosemide.

The goal of the ATC differs from the one of the FTC, yet the two resources are sharing some very similar concepts, which can be used for the evaluation. Categories of both classifications contain approved drugs with a DrugBank iden-

tifier, meaning that some of the drugs indexed in the FTC are also present in the ATC. From that, it is possible to define some evaluation points, which will help to assess the automated classification process.

3.2.6.1 Evaluation Points

An evaluation point is defined as an equivalence between a class from the FTC with one or more classes from the ATC. The idea is to look at the set of drugs contained in both side of the equivalence and estimate the overlap, as illustrated in the Figure 3.8.

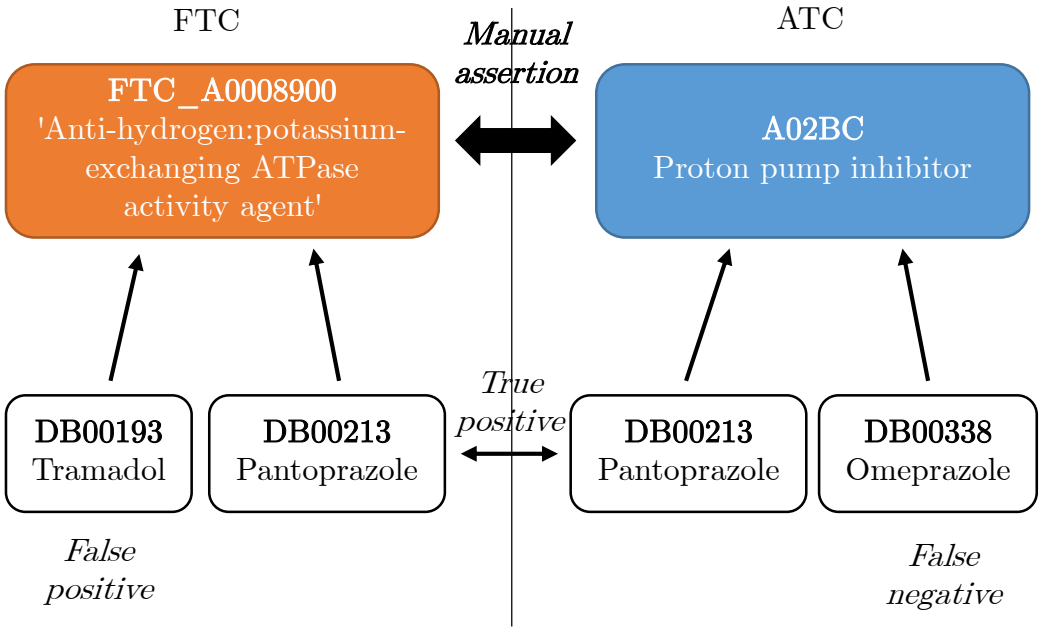


Figure 3.8: Example of evaluation point. An manual assertion is made between an ATC category (blue) and a FTC class (orange) when the two concepts are semantically equivalent. Then drugs belonging to each of these classes are compared, and the evaluation can be performed.

Evaluation points are defined by hand and not themselves evaluated. The full list of evaluation points as well as a summary of the results are available online at `https://www.ebi.ac.uk/chembl/ftc/evaluation/`. Each evaluation point has a series of true/false positive and false negative drugs associated with it.

3.2.6.2 True Positives

Drugs that are present in both the FTC and the equivalent ATC class(es) are called true positives. These compounds reflect that the automated classification was capable of retrieving correctly the information present in the gold standard (ATC).

3.2.6.3 False Negatives

These drugs are present in the ATC class(es) but not in the corresponding FTC class. The automated classification failed to retrieve these compounds. The smaller the number of false negatives is, the better the FTC is at recalling drugs. A small number of false negatives means that if a drug is present in the ATC (gold standard), then it is likely that the drug will also be correctly categorised in the FTC.

3.2.6.4 False Positives

The false positives are the drugs present in the FTC category of the evaluation point but not in the corresponding ATC classes. A high number of false positives means that the FTC is "over-assigning" compounds to classes. The false positives relates to the accuracy of the classification. In the context of this work, some false positives could also be considered as drug repositioning opportunities.

3.2.6.5 Precision

Recall is the probability that a randomly selected drug from the ATC has been assigned to the correct corresponding class in the FTC. The value is standardised as a percentage and corresponds to the formula: $TruePositive/(TruePositive + FalseNegative)$.

3.2.6.6 Recall

Recall is the probability that a randomly selected drug from the ATC has been assigned to the correct corresponding class in the FTC. The value is standardised as a percentage and corresponds to the formula: $TruePositive/(TruePositive + FalseNegative)$.

3.2.7 Semantic similarity

The semantic similarity measure performed over the FTC is a derivative of the Jaccard index (Jaccard, 1912) (Rogers and Tanimoto, 1960) (see Chapter 2, section 2.4.3). It is probably best understood as an example: if two classes A and B are considered, the semantic similarity between these classes corresponds to the number of OWL superclasses (direct and indirect, obtained with a reasoner) that are shared by A and B (intersection) divided by the number of superclasses of A or B (union). The index ranges from 0 (totally different) to 1 (identical). A similar approach was successfully implemented by (Hoehndorf et al., 2011b), for similarity computations over phenotypic traits.

3.2.8 Mode of action similarity against indication

A statistical analysis was performed over the data presented on section 3.5.2. When two compounds are randomly taken, they have on average a higher mode of action similarity when they are assigned to the same ATC category (one ATC level). In order to estimate whether this observation was due to chance only, I formulated the following null hypothesis (H0): *for a pair of drug A and B, it does not matter to which ATC category they belong to, their similarity is always average.* The alternative hypothesis (H1) was: *for a pair of drug A and B, it matters in which ATC category they belong to, their similarity is significantly different than average.*

A permutation test was then performed for each ATC category. For example, I started with the ATC category A (first row on Figure 3.12), looked at the similarity values when pairs of compounds both belong to the category A (top right corner square) and compared it to the similarity values when the pair of compounds belong to different categories (A and B, A and C and so forth). For each comparison I obtained two distributions of values (not shown). On average the similarity values are always higher when the two compounds belong to the same category (A/A versus A/B for instance). A permutation test (n = 20,000) was performed in order to see whether this observation was due to chance only. I was able to reject the null hypothesis for a significance level of 0.05 all the times. The choice for a permutation test was driven by the fact that MoA similarity values do not follow any type of standard distribution (data not shown).

3.2.9 Knowledge base specification

This section presents the scaffold of the knowledge base underlying the FTC. The logic structuring the FTC comes essentially from a set of core OWL properties (rich RBox). Some of these properties originate from the GO. When necessary some new ones have also been introduced. In order to understand how these properties interact, first will be presented the fundamental classes present at the top of the FTC classification, followed by the presentation of the object properties.

3.2.9.1 Core FTC classes

molecular function

- Identifier: `http://purl.obolibrary.org/obo/GO_0003674`
- Label: 'molecular function'
- Definition: As defined by the Gene Ontology: Elemental activities, such as catalysis or binding, describing the actions of a gene product at the molecular level. A given gene product may exhibit one or more molecular functions.

biological process

- Identifier: `http://purl.obolibrary.org/obo/GO_0008150`
- Label: 'biological process'
- Definition: As defined by the Gene Ontology: Any process specifically pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms. A process is a collection of molecular events with a defined beginning and end.

Protein

- Identifier: `http://purl.uniprot.org/core/Protein`
- Label: 'protein'
- Definition: As defined by Uniprot: Description of a protein.

- Comment: Gene products present inside the FTC are all human proteins. Uniprot URIs are used.

Drug

- Identifier: `http://schema.org/Drug`
- Label: 'drug'
- Definition: As defined by schema.org: A chemical or biologic substance, used as a medical therapy, that has a physiological effect on an organism.
- Comment: In the context of the FTC, DrugBank chemicals are considered for their role as therapeutic agent rather than for their chemical structure.

therapeutic agent

- Identifier: `https://www.ebi.ac.uk/chembl/ftc/FTC_C1`
- Label: 'therapeutic agent'
- Definition: Role of a drug capable of producing a therapeutic effect.

DrugBank compound

- Identifier: `https://www.ebi.ac.uk/chembl/ftc/FTC_C2`
- Label: 'DrugBank compound'
- Definition: Drug coming from DrugBank.

3.2.9.2 Core FTC properties

part-of

- Identifier: `http://purl.obolibrary.org/obo/BFO_0000050`
- Characteristic: Transitive
- Label: 'part-of'
- Definition: As defined and used in the Gene Ontology. More information at `http://www.geneontology.org/GO.ontology.relations.shtml#partof`

has-part

- Identifier: `http://purl.obolibrary.org/obo/BFO_0000051`
- Characteristic: Transitive
- Label: 'has-part'
- Definition: As defined and used in the Gene Ontology. More information at `http://www.geneontology.org/GO.ontology-ext.relations.shtml#haspart`

regulates

- Identifier: `http://purl.obolibrary.org/obo/RO_0002211`
- Chained property: 'regulates' o 'part-of' $\rightarrow$ 'regulates'
- Label: 'regulates'
- Definition: As defined and used in the Gene Ontology. More information at `http://www.geneontology.org/GO.ontology.relations.shtml#regulates`

negatively-regulates

- Identifier: `http://purl.obolibrary.org/obo/RO_0002212`
- subPropertyOf: 'regulates'
- Label: 'negatively-regulates'
- Definition: As defined and used in the Gene Ontology. More information at `http://www.geneontology.org/GO.ontology.relations.shtml#regulates`

positively-regulates

- Identifier: `http://purl.obolibrary.org/obo/RO_0002213`
- subPropertyOf: 'regulates'
- Label: 'positively-regulates'

- Definition: As defined and used in the Gene Ontology. More information at `http://www.geneontology.org/GO.ontology.relations.shtml#regulates`

involved-in

- Identifier: `https://www.ebi.ac.uk/chembl/ftc/FTC_R1`
- Label: 'involved-in'
- Domain: 'protein'
- Range: 'biological process'
- Definition: Entails the participation of a protein in a biological process

has-function

- Identifier: `https://www.ebi.ac.uk/chembl/ftc/FTC_R2`
- Label: 'has-function'
- Domain: 'protein'
- Range: 'molecular function'
- Definition: Describes the molecular function born by a protein

perturbs

- Identifier: `https://www.ebi.ac.uk/chembl/ftc/FTC_R3`
- Label: 'perturbs'
- Domain: 'drug'
- Range: 'protein'
- Definition: Specific biochemical interaction through which a drug substance will affect the activity of a protein. The property refers to the specific molecular targets to which the drug binds, such as an enzyme or receptor.

negatively-perturbs

- Identifier: `https://www.ebi.ac.uk/chembl/ftc/FTC_R4`
- Label: 'negatively-perturbs'
- subPropertyOf: 'perturbs'
- Definition: Specific biochemical interaction through which a drug substance will decrease the activity of a protein. The property refers to the specific molecular targets to which the drug binds, such as an enzyme or receptor.

postively-perturbs

- Identifier: `https://www.ebi.ac.uk/chembl/ftc/FTC_R5`
- Label: 'postively-perturbs'
- subPropertyOf: 'perturbs'
- Definition: Specific biochemical interaction through which a drug substance will increase the activity of a protein. The property refers to the specific molecular targets to which the drug binds, such as an enzyme or receptor.

3.3 The classification

The knowledge base behind the FTC is built by integrating information coming from various sources. The GO terms serve as template to create the FTC categories describing the MoAs; DrugBank provides the known links between drugs and their protein targets and Uniprot maps targets to their respective GO annotations. Drugs are further assigned into MoA categories according to the OWL constructs and axioms defined in the FTC. A reasoner, a program capable of understanding such axioms, performed this task (see the method section 3.2). The process to build the FTC is summarised in Figure 3.9 alongside an example.

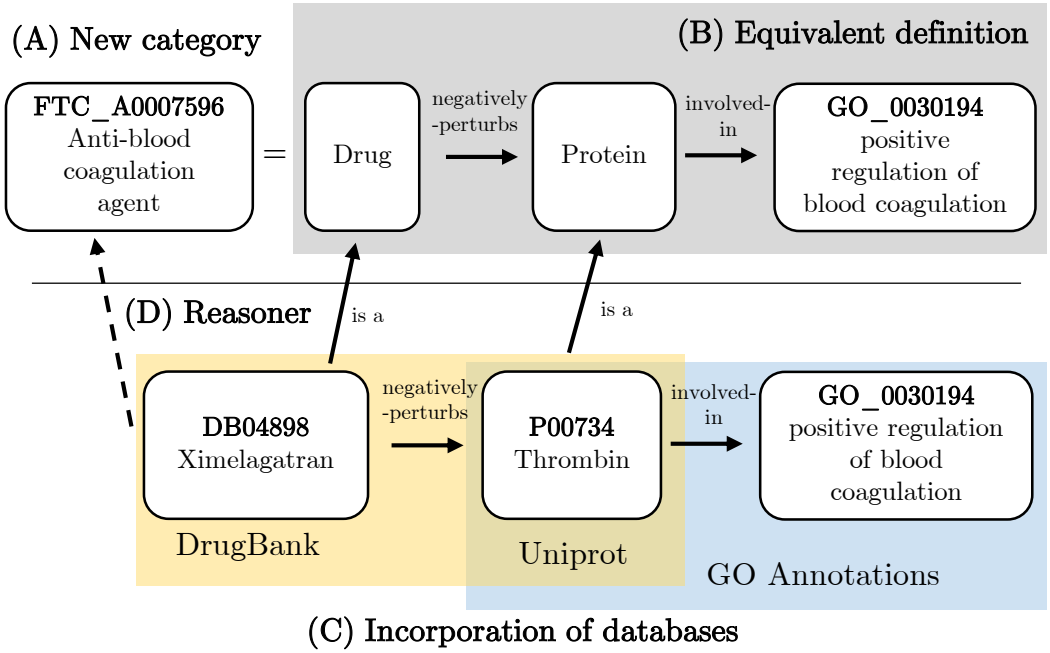


Figure 3.9: The diagram gives an overview of the integrated resources and building process. (A) The name of FTC categories representing MoAs are directly derived from the GO terms representing the molecular functions and biological processes. (B) Each of the new FTC class has a logical equivalent definition assigned to it (axiom), representing the necessary and sufficient conditions for a drug to be classified in the corresponding MoA class. (C) The content of various databases is incorporated and linked using the FTC specific logical properties. (D) Finally a reasoner classifies the knowledge base and assigns drugs to MoA classes based on whether or not a definition can be satisfied. For example, the drug *ximelagatran* will be assigned as member of the category *Anti-blood coagulation agent* because of the logical links *ximelagatran negatively-perturbs prothrombin* and *prothrombin involved-in positive regulation of blood coagulation*. The taxonomic structure of the FTC appears also in the reasoning step, from the entailment of the equivalent definitions.

The core step is the generation of axiomatic representations of MoAs by decomposing GO types into positive and negative regulations of biomolecular functions and processes. The help of reasoning techniques we can further derive and assign MoA across the knowledge base to given drugs. It requires a few seconds (four processing cores, 8 GB RAM) to classify the knowledge base (ELK reasoner). Other OWL reasoners (e.g, Hermit, Pellet, etc.) disqualified mainly due to long processing time (data not shown - see Gonçalves et al. (2013) for time values).

The FTC forms a taxonomic structure as illustrated on Figure 3.10, which

arises when the reasoner classifies the knowledge base. In general, categories may have multiple parents and multiple children (see `https://www.ebi.ac.uk/chembl/ftc` for interactive use).

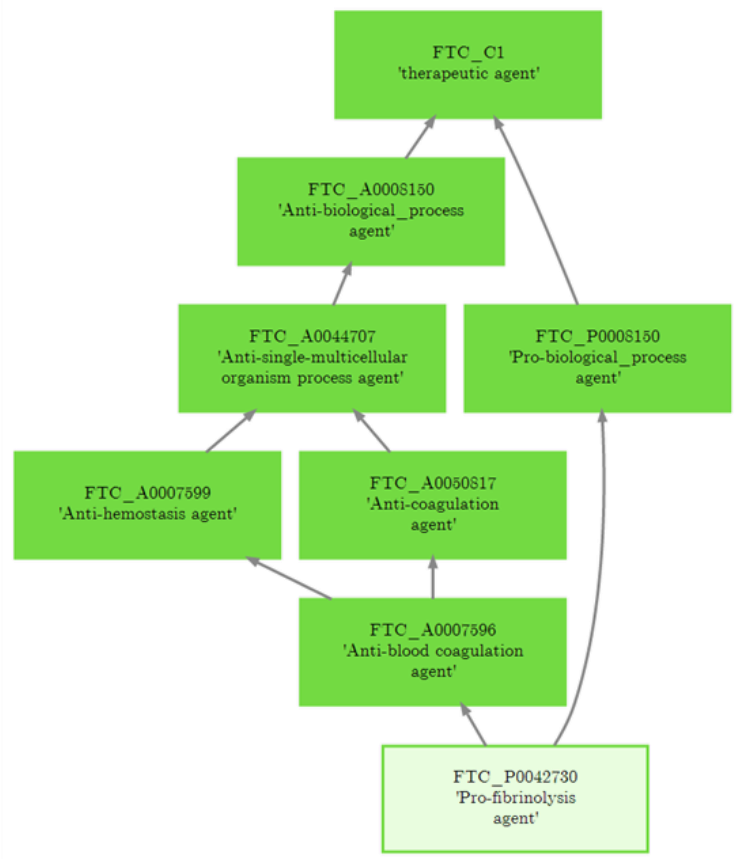


Figure 3.10: Parent categories to the FTC class *Pro-fibrinolysis agent* (FTC_P0042730). The classification is a direct acyclic graph where categories are describing increasingly specific concepts. Arrows entail subclass relationships between the terms (*is a* relation).

In total there are 1,280 FDA-approved DrugBank compounds (chemical and biotherapeutics) associated with 1,264 human protein targets, where each drug is acting on at least one human protein target. The FTC introduces 23,353 new categories describing the mode and mechanism of action of therapeutic compounds. 4,289 of these categories belong to the biological processes in GO and 19,064 to the molecular functions. A summary of the metrics behind the latest build is available online at `https://www.ebi.ac.uk/chembl/ftc/evaluation/`. Out of all FTC categories, 1,432 categories (>6%) directly contain at least one approved drug. This number increases up to 2,532 (>11%) when direct and indirect drugs

are considered. FTC categories not containing drugs (e.g, FTC_A0001771 - *Anti-immunological synapse formation agent*) represent MoAs for which no approved compounds has qualified yet or that have not been identified as such in the FTC.

3.4 Evaluation

The content of the FTC has been evaluated against the drug categorisation of the ATC, which has been produced by manual curation and serves as a gold standard. *A priori*, both resources serve different purposes and as a consequence, the evaluation has to take this into consideration (cf section 3.6.2). The full methodology behind the evaluation is described in the section 3.2.6 of the methodology. Briefly, for 68 categories from the FTC one can manually identify a set of semantically equivalent categories in the ATC. I call these equivalent categories the *evaluation points*. All drugs from each evaluation point were then assessed to determine the quality of the FTC against the gold standard, i.e. the ATC. For example, the FTC category *Anti-hydrogen:potassium-exchanging ATPase activity agent* (FTC_A0008900) has been manually asserted as equivalent to the ATC category *Proton pump inhibitors* (A02BC). A summary of this evaluation point is furthermore available online at `https://www.ebi.ac.uk/chembl/ftc/evaluation/FTC_A0008900`.

For 1,280 DrugBank compounds in the FTC, 1,134 are also present in the ATC, therefore only those were considered. The *evaluation points* cover a total of 471 DrugBank compounds or around 41% of common drugs to both classifications. Out of these, 275 compounds are true positives, i.e. they match both, the FTC and ATC categories for a given evaluation point. The *proton pump inhibitor* evaluation point is such a case where all the drugs (omeprazole, esomeprazole, pantoprazole, lansoprazole, rabeprazole) are present both in the FTC category and in the corresponding ATC category. The total number of compounds from an ATC category but where it was not possible to identify a corresponding FTC category is 35 (false negatives). Finally, 280 compounds are present in a FTC class but not in any corresponding ATC category (false positives). Overall a recall of 89% is derived; this percentage indicates that the automatic build of the FTC covers a good portion of the content already present in the ATC. The precision of 50% shows that the FTC contains for a given MoA many more drugs

than the equivalent ATC categories. This result was expected and comes the original idea behind the FTC: representing in a systematic fashion the implicit and explicit MoAs of drugs, in particular the ones not already indexed by current classification scheme.

3.5 Exploration

The FTC was designed to assist drug repositioning analyses by explicitly representing the polypharmacology of approved drugs. In this section I exemplify how the resource can be used to perform different types of analysis, whose will be extended in Chapter 4.

3.5.1 Polypharmacology spectrum

The more information on a drug's molecular targets and their physiological roles, the more opportunities exist to re-orient a drug into doing something new. The therapeutic agents described in the FTC can have several MoAs, i.e. may be acting on different biomolecular functions or processes, which demonstrates the intrinsic polypharmacology of the approved compounds. Figure 3.11 illustrates the polypharmacology spectrum by showing the distribution of number of MoAs per compound.

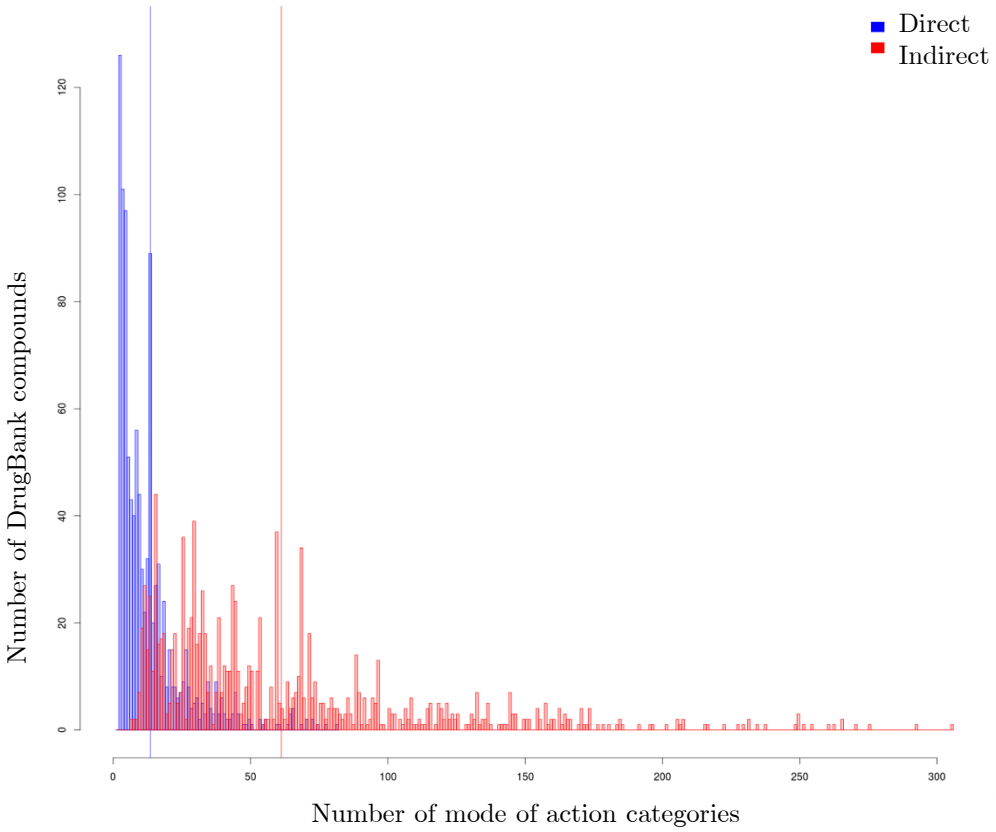


Figure 3.11: Distribution of the direct (blue) and indirect (red) number of MoAs per drug. Means are indicated with a solid line. On average each compound has 13.5 MoAs when only direct classes are considered. The number rises to 61.2 when indirect MoAs are included. Indirect MoAs are the ancestor classes in the taxonomy as shown in Figure 3.10. The distribution range is wider when indirect MoAs are considered (range=299; min=7; max=306) versus direct MoAs only (range=79; min=3; max=82). These results emphasises the fact that some drugs are well characterised in databases and could be used for a variety of specific biological tasks. Finally some compounds have been assigned to a small number of FTC categories; in such cases little is known or reported about their pharmacology and repurposing opportunities might be limited.

When only direct categories are considered, compounds belong on average to 13.5 MoA categories. This number increases to 61.2 when parent categories are taken into consideration (super classes). Not all the MoAs are relevant to a disease, some FTC categories are particularly abstract (e.g, *Anti-biological process agent*) yet they represent discrete categories to which the drug belongs with an explicit and clear meaning. These discrete MoAs are a good starting point to understand what a compound can do when administered in a human system.

Compound's polypharmacology is well represented in the FTC, as shown by the numerous MoAs each approved drug can exhibit.

I decided to further look at a well-known repositioning example, in order to see whether the FTC was suitable to identify the new uses of an old drug. I picked thalidomide for this exercise (`https://www.ebi.ac.uk/chembl/ftc/agent/DB01041`). The molecule was first indicated to treat morning sickness in pregnant women, but has been quickly abandoned after its developmental toxicity has been discovered in newborns (see section 1.2.2). The accepted molecular mechanism behind the side-effect is an impairment of the angiogenic process responsible for the development of members, affecting in particular the limbs (Therapontos et al., 2009). I found that thalidomide was accurately classified as *Anti-cell migration involved in sprouting angiogenesis agent* in the FTC, capturing the known toxicity of the drug. Furthermore, thalidomide currently investigated for a multitude of new usages, in particular for anti-cancer and immunomodulatory activities among others (Teo et al. (2005) and section 1.2.2). These new indications are well represented in the FTC too, for example by the categories *Anti-vascular endothelial growth factor production agent* or *Anti-cell division agent* for antineoplastic activities, or by the classes *Anti-cytokine secretion agent* and *Anti-I-kappaB kinase/NF-kappaB cascade agent* for its effect on the immune system. These observations demonstrate that the FTC can successfully capture the molecular reasons behind the repositioning of an old compound, relying on automated reasoning over integrated electronic evidence. Moreover the classification can also provide valuable insight regarding potential toxicity too.

3.5.2 Drugs with similar functions have similar indications

The list of MoAs attributed to a drug can be exploited as a descriptor for the therapeutic agent: the tree structure of the FTC can be used to derive some similarity metrics over the MoAs. The underlying heuristic is to assume that the closer two entities are in the taxonomy, the more similar they are. I used a straightforward approach derived from the Jaccard index (see section 3.2.7) in order to compare approved drugs based on the similarity of their MoAs. For instance, the similarity between two compounds present in the same FTC category is 1 (maximum). The similarity between an *anti-blood coagulant* and *pro-blood*

coagulant is 0.29, reflecting the fact that such compounds are dissimilar with regards to the outcome of their biological effect. As the MoA is intuitively expected to be the central concept leading to the indication of the drug, I expected that on average, drugs with similar MoAs would be indicated towards similar therapeutic areas. The heat map presented in Figure 3.12 shows a pairwise comparison of all the drugs of the FTC based on their relative MoA similarity.

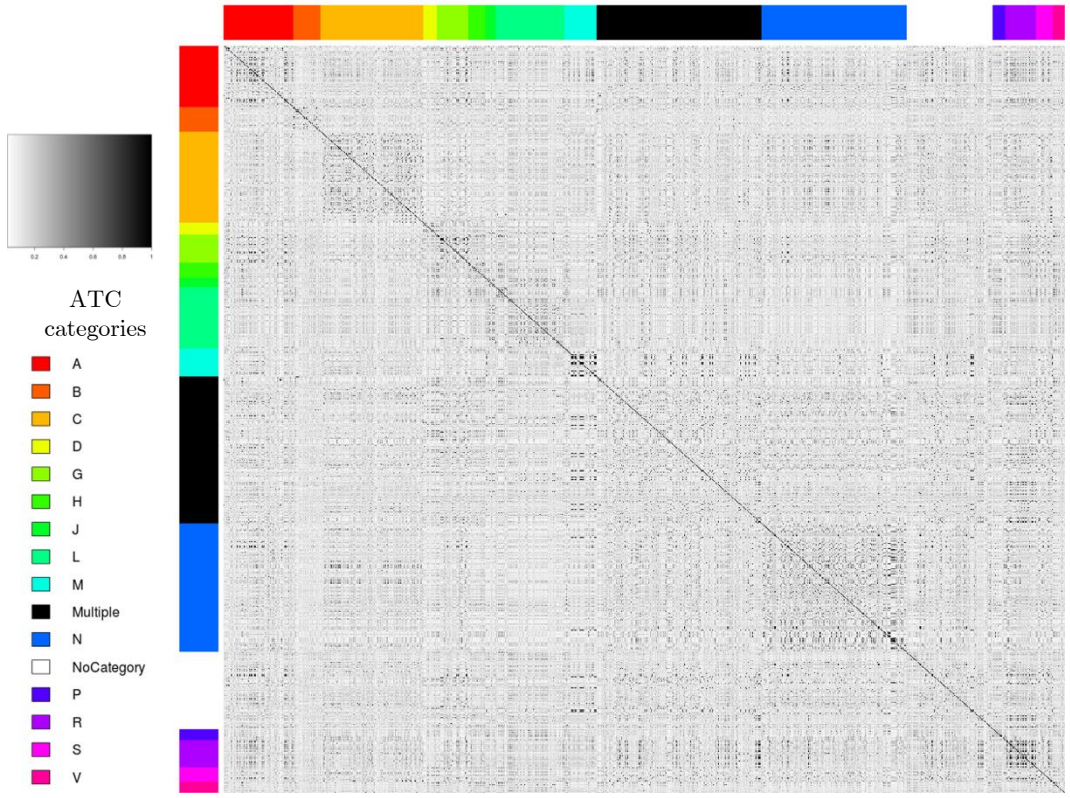


Figure 3.12: Pairwise comparison of MoAs similarities. Therapeutic indications are represented by ATC categories which are the colours on the side. For instance, the compound reteplase (DB00015) has the ATC code B01AD07, which appears as *B* (dark orange) on the plot. Only the first ATC level is considered. The similarity descriptor ranges from 0 (not similar - white) to 1 (identical - black). Some compounds belong to multiple ATC categories (*Multiple*) and some others do not have an ATC code (*NoCategory*). The average similarity of drugs present in the same therapeutic category is significantly higher on average when separately compared to all other indications.

The compounds are further grouped by therapeutic indications as defined by the ATC. The heat map reveals some square patches around the central diagonal; the overall similarity appears higher when compounds from the same ATC group

are considered. A significance analysis (see method, section 3.2.7) revealed that the average MoA similarity of compounds belonging to the same ATC category is significantly higher than when compounds belonging to different categories are compared. Indeed, for each category, the p-value was inferior to 0.05 based on 20,000 random permutations over the similarity values. This result supports the idea that drugs with similar MoAs have similar indications. Note that the mean of the similarity values was considered for the statistical analysis; some outliers are also present in the map, which can be interpreted as repositioning hypotheses. These outliers have indeed similar MoAs, yet they belong to totally different therapeutic areas and are used for different purposes according to the ATC. Such cases will be further analysed and discussed in Chapter 4. Hypotheses have to be manually examined and interpreted, as ATC categories are only covering some of the legal usage of the drugs. I expect to find off-label indications in the predictions for instance, as well as some false positives.

Figure 3.13 present similar association behaviour when two levels of the ATC are considered (no statistical significance performed). In this case the higher intensity squares are smaller, reflecting the finer resolution of the therapeutic areas (2 ATC levels).

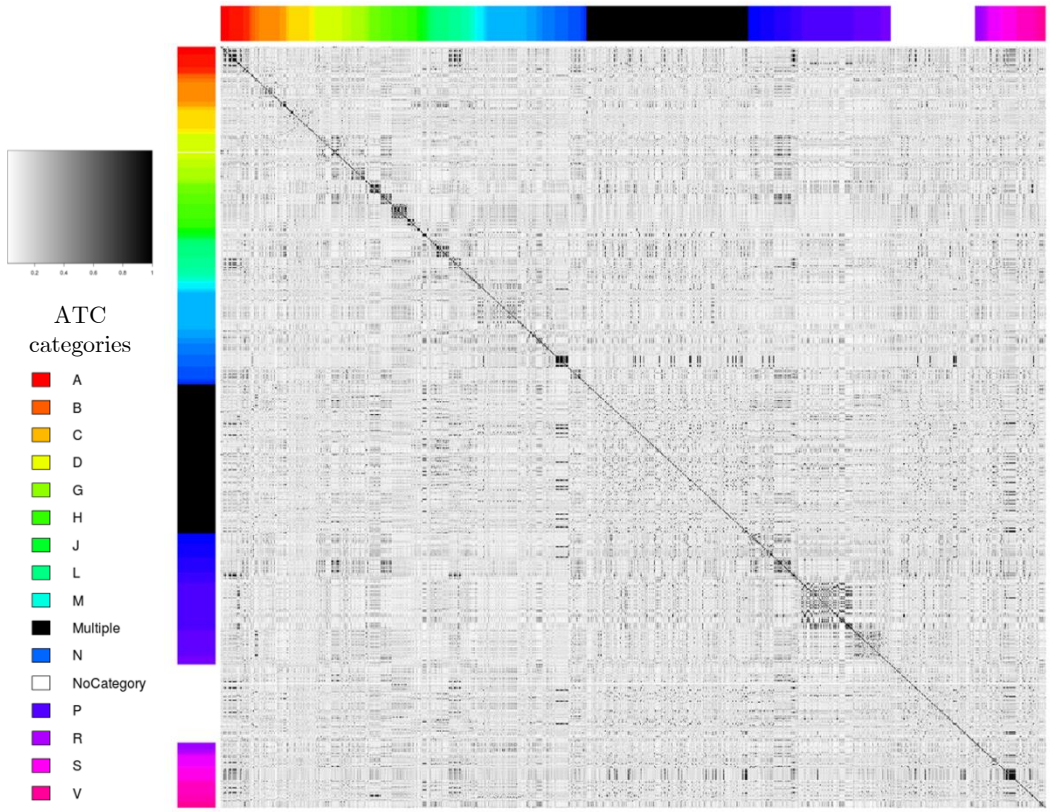


Figure 3.13: Pairwise comparison of MoAs similarities. Therapeutic indications are represented by ATC categories which are the colours on the side. Two ATC level are considered on this graph, as opposite to Figure 3.12, where only one level was considered. This increased resolution allows to identify more granular square patterns along the diagonal, where drugs from the same groups appear to have higher intensity values (no analyses performed).

Figure 3.14 re-uses the same data as Figure 3.12 (one ATC level) but with a clustering function apply to it (hierarchical clustering - Manhattan distance) in order to reveal functional clusters of drugs (no analyses performed on this processing). Noteworthy, the taxonomic tree generated on the top of the data reflects the structure of the FTC.

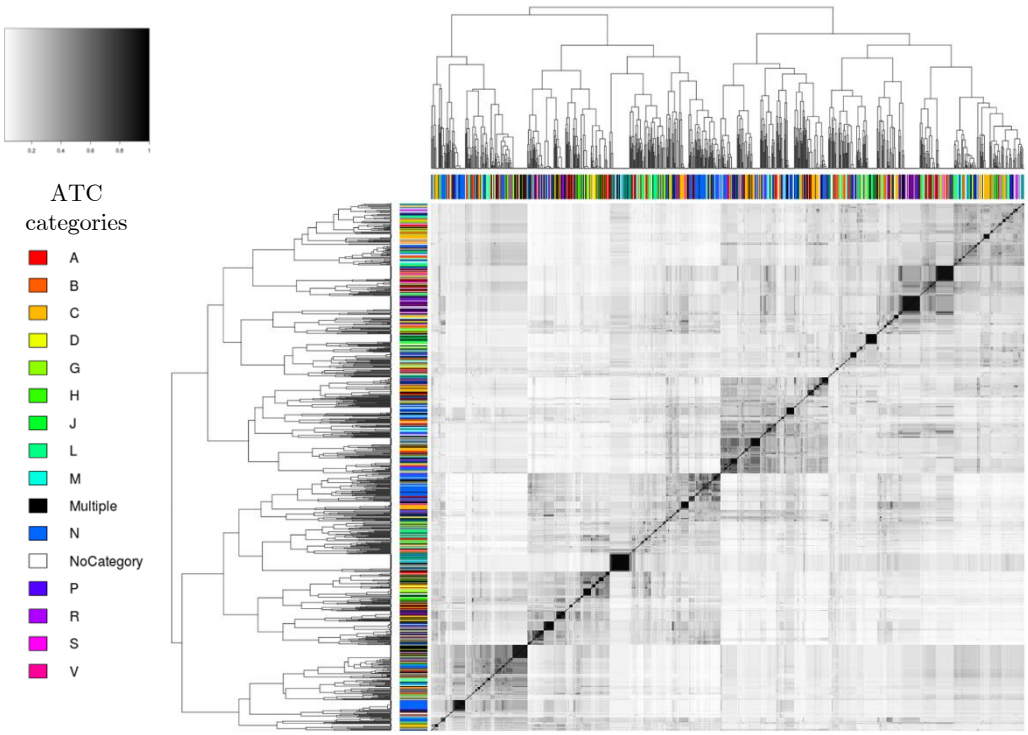


Figure 3.14: Pairwise comparison of MoAs similarities. Therapeutic indications are represented by ATC categories which are the colours on the side. One ATC level is considered on this graph, just as in Figure 3.12. The similarity values are further clustered hierarchical clustering based on the Manhattan distance. This processing of the data enables to see functional clusters of drugs, namely groups of drugs with a similar pharmacology (no analyses presented in this document).

Finally, Figure 3.15 shows the distribution when compounds are sorted based on their identifiers; no patterns are identifiable in this case. Figure 3.15 acts as a visual control, reflecting the result obtained from the similarity analysis.

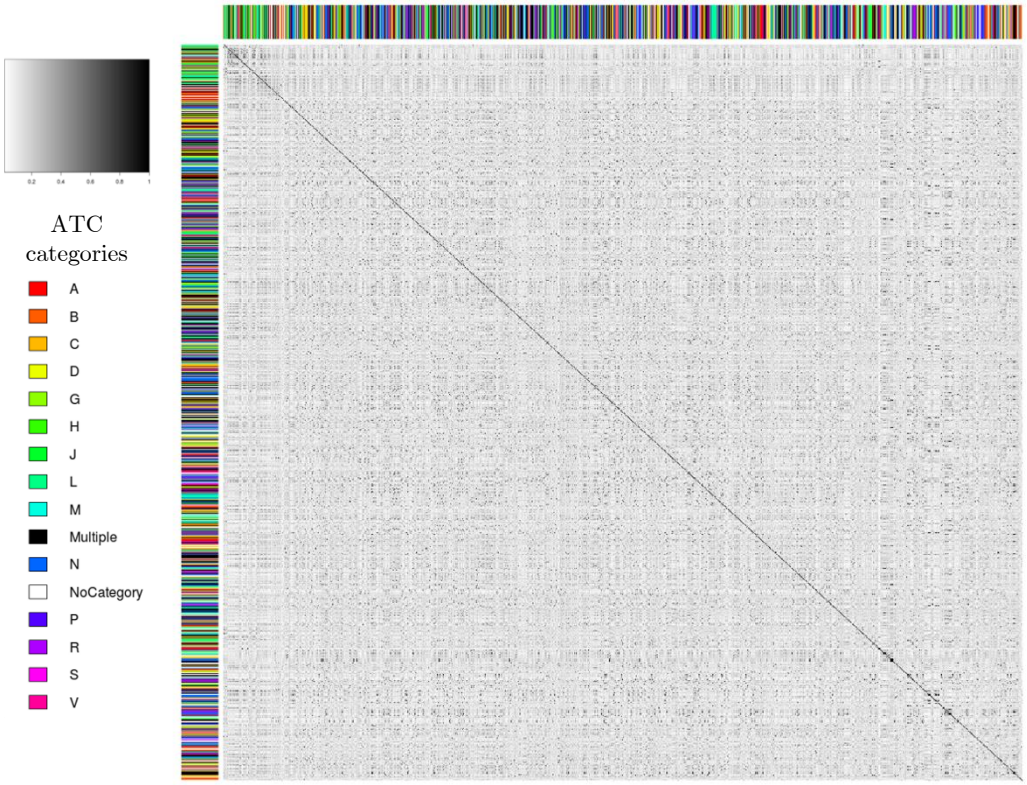


Figure 3.15: Control pairwise comparison of MoAs similarities. Therapeutic indications are represented by ATC categories which are the colours on the side. One ATC level is considered on this graph, just as Figure 3.12. Drugs are randomly sorted (yet symmetrically). No visual pattern is observable in this case, as opposed to what is seen on Figures 3.12, 3.13 and 3.14.

Taken together, these results emphasise that the MoAs as defined in the FTC are indeed on average associated with the therapeutic indication of a drug. This result supports the validity of the resource and its potential to computationally address indication discovery. Further analyses are presented in Chapter 4, also integrating the concept of *chemical structure*.

3.6 Discussion

The FTC is a novel classification for approved drugs, which can be used as a starting point to generate drug repurposing hypotheses. This classification leverages the information present in various databases and ontologies, similarly to

the Open PHACTS initiative (Williams et al., 2012) and to the work done by Hoehndorf et al. (2012) and Jupp et al. (2012). The FTC mostly differentiates itself from these projects by providing a whole set of new categories on the top of the integrated information, dedicated to tackle a very specific problem: drug repositioning. Moreover, the semantic model behind the integration is richer than any of the previous approaches: FTC properties are expressive, thanks to the use extensive of transitive or chained property axioms (see section 3.2.9).

3.6.1 Biological assumptions

An asset of the FTC is its ability to handle efficiently categorical data: classes and relationships are accurately defined, in order to classify compounds based on the semantics of their relations. The properties linking drugs to their respective protein targets (*positive* and *negative perturbations*) are however simplistic. At the time being, no consideration is given regarding the binding strength between the drug and the proteins, yet it is a key factor to derive potent and specific activities in the human body. This is also the case for other types of numerical data, such as the dosage; the FTC can predict a role for a drug, yet it cannot provide any information about the concentration or the administration route necessary to obtain the potential effects. The current relations between targets and their involvement in biological processes are also not a fully accurate representation of the biological phenomenon. In a cell, specific domains of the protein could mediate different functions. Only one of such activity types can sometimes be inhibited by a drug (Kruger et al., 2012), yet I am assuming in the FTC that as long as a drug affects a protein, it can therefore alter all its known functions. These limitations come from the semantics behind the axioms structuring the classification, themselves based on the information available from the databases.

Despite entailing not entirely accurately the biochemical reality, the axioms help to generate a larger number of hypotheses, the primary goal of the FTC and in-line with the theory presented on Chapter 2. The dosage issue is partially addressed by the *regulator pattern* (see section 3.2.2): it should be easier to experimentally adjust the concentration of the compounds classified as *pro* or *anti* biological process agents in order to modulate a physiological effect.

The predictions generated by the FTC depend on the resolution of the curated information released by the original data providers. Erroneous or missing

information will lead to misclassification by the reasoner. Some expected outcomes are also missing from the predictions; sildenafil for instance was expected to be classified as *pro-penile erection agent* (FTC_A0043084), yet the lack of appropriate GO annotation prevents it. After discussion with the GOA curation team, a manual annotation can only be asserted based on published experimental results. No document was found to support the involvement of the cGMP-specific 3',5'-cyclic phosphodiesterase (sildenafil's main target) in the *negative regulation of penile erection* (GO:0060407), therefore no annotation can be made. Further work could be done in this direction, by trying to automatically infer more annotations or by using the electronically generated ones, in order to generate broader yet potentially less plausible repositioning hypotheses.

3.6.2 Interpreting the evaluation

Out of the evaluation, the high recall value (89%) supports the idea behind the automated build of the FTC: the data from different repositories funded and curated in parallel, can be integrated to automatically create a new resource. This new classification (FTC) contains most of the known information present in an external gold standard (ATC) and relies on DLs to leverage the native information. In the context of this work I compared the content of the FTC against the ATC, knowing that these two taxonomies have diverging goals. During the evaluation, equivalences have been manually asserted between classes, which are assumed to have fairly similar meaning and containing similar sets of compounds. These manual assertions are however a weakness, as they are themselves not evaluated (free parameter). The presence or absence of a link was determined only by one curator and any mistake can influence consequently the recall and precision values. The precision of 50% tells that the FTC tends to over-assigns compounds to MoA categories. The low precision value is acceptable in this case, as one of the underlying motivation of the FTC is to broadly represent polypharmacology, specially the one not present in gold standards such as the ATC, referencing only regulated usage. In this regard, the evaluation should be considered more as a safety control rather than a formal assessment of a predictive method.

The false positives derived from the evaluation can also be considered as drug repurposing hypotheses: these drugs can indeed be interpreted as suitable for the ATC category, yet not indexed as such. However, these predictions should be in-

terpreted with caution, as it is currently impossible to distinguish a false positive from a reprofiling opportunity. These considerations do not interfere with the exploration based on semantic similarities. Finally, note that the ATC/FTC equivalences are open and editable online (`https://github.com/loopasam/ftc/blob/master/data/drugbank-relations-mappings.txt`), any modification will be automatically incorporated in the next release of the resource. It is also possible to evaluate the FTC against a different taxonomy, like the Medical Subject Headings for example, which can be subject to future work.

3.7 Summary

The representation of the MoA as motivated in Chapter 1 was presented in this chapter. The axioms behind the resource and the deductions reached by the reasoner follow the theory introduced in Chapter 2. Despite the large size of the knowledge base, the classification process is fast and scalable, thanks to the $\mathcal{EL}^{++}$ profile. I will further present an analysis of the relationship between the MoA, the indication of a drug and its molecular structure in the coming chapter. Drug repositioning use-cases, not presented in the section, will also be discussed.

To conclude, the FTC is public resource, which can assist drug repositioning initiatives or enhance computational studies that evaluate drugs according to their *mode of action*. The resource attributes biomolecular functions and processes to drugs, the same way as GO types have been assigned to gene products; its role is analogous to the one of a toolbox, classifying items based on their use (see Figure 3.16).

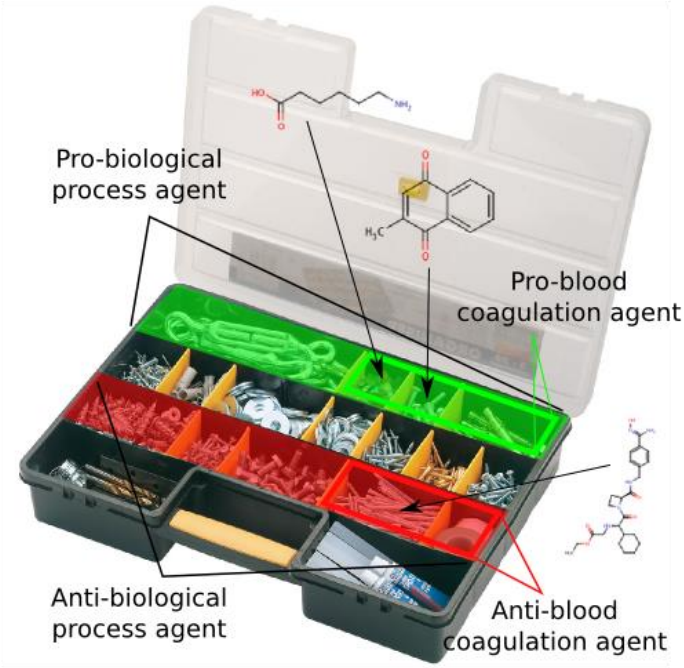


Figure 3.16: Pharmacological toolbox analogy. The FTC describes categories inside which drugs can be classified; the classification helps to select the right tool for the right task, similarly to a toolbox.

The construction of FTC relies on an axiomatic representation as the core means to attribute and derive the MoA for approved drugs. I showed the validity of the approach by comparing the content of the FTC to a well established clinical gold standard, the ATC. In Chapter 4, I analyse the FTC's content to illustrate how repositioning hypotheses can be generated for hypertension treatment and Alzheimer's disease, using two different methodologies.

CHAPTER 4

Systematic drug repositioning analysis

Key points

- The Functional Therapeutic Chemical Classification System (FTC) provides a descriptor for the function of drugs, which can be used to analyse the systematic relationship between function, structure and indication for approved drugs.

- Overall, approved drugs have dissimilar structures and dissimilar functions. When the therapeutic area is considered, the more specific the drug's indication is, the more similar the structure and function are. These observations are in line with the *similar property principle* and can be used to isolate drug repositioning hypotheses from the dataset considered.

- A drug repositioning hypothesis is a pair of drugs indicated for different therapeutic areas (two ATC levels) but with a high functional similarity value. This method of extraction is called *similarity-based*. I extracted 797 of such pairs and made them openly available through a web application.

- The set of repositioning hypotheses can be used to study the relationship between therapeutic areas: some repositioning hypotheses are more frequent between particular therapeutic groups and overlap with known off-label uses.

- From the full set of hypotheses, two use-cases are investigated in detail: hypertension and Alzheimer's disease. The drug *similarity-based* repositioning hypotheses are compared against the *toolbox approach*, where FTC categories are directly used as a starting point to forward repurposing opportunities.

- Cardiovascular hypertension hypotheses relate mostly to similar pharmacological actions but in different anatomical areas. Alzheimer's disease hypotheses involve biological processes related to the Tau protein and beta-amyloids; the majority of hypotheses find supporting evidence in the biomedical literature.

- This analysis concludes the thesis. Chapter 5 sets the work done in the context of the overall drug discovery process and discusses future work and open questions.

Author's comment

This chapter is a narrative analysis of the content of the FTC. Results and discussion are mixed together as I believe it makes the text more understandable and clarifies the reasoning behind the work. A methods section is provided at the end, describing the details of the analyses performed.

4.1 Introduction

I decided to look at drug repositioning from the perspective of the mode and mechanism of action (Chapter 1). The implementation of a solution required first some theoretical basis to be set, namely the black box model of the cell, specified using description logics (Chapter 2). The FTC then implemented the theory, as presented in the previous chapter (Chapter 3), and was evaluated against existing solutions to assess the relevance of the approach. This coming chapter will present the biological analysis performed over the content of FTC, in order to more directly address drug repositioning concerns.

As the FTC characterises the concept of function in a systematic fashion, the obvious questions to ask are: how does it relates to the chemical structure? Do

similar compounds have similar functions? What about the indication? Is there any relation between function, structure and therapeutic usage?

In this regard, the analysis below first identifies the repositioning hypotheses present in the FTC and discusses the relevance of function in the process. Secondly, the extracted hypotheses are examined in depth and interpreted from a biological viewpoint. Because the FTC provides systemic insight on the drug repositioning topic, it is therefore possible to explore the broad relationship between therapeutic areas as well as the connection between drug repositioning and off-label uses. Finally, a couple of detailed use-cases are discussed: cardiovascular hypertension and Alzheimer's disease. These pathologies will serve to demonstrate how two different methodologies can be applied over the FTC to extract hypotheses: the *similarity-based* and the *toolbox* approaches.

This chapter is more focused on the biology and its interpretation. The content of FTC is dissected using semantic similarity and mode of action (MoA) cherry-picking. The conclusions drawn in this chapter will introduce future work to be done, as well as new leads to be explored.

4.2 Structure, function and indication of drugs

The most commonly accepted rule in drug discovery is probably the *similar property principle*: similar structures have similar biological activities - or functions (Martin et al., 2002) (Kubinyi, 1998) (Johnson and Maggiora, 1990). Despite being true in some cases, there are plenty of examples in contradiction with this rule. As the FTC provides a unique characterisation of the function of chemical compounds, I decided to analyse the structure/function relationship for approved drugs, with a systematic stance. Before being able to do so, it is however necessary to choose adequate descriptors; the function will be represented by the FTC, but the structural descriptor still remains to be selected. This preliminary characterisation will allow me to then compare each drug's chemical structure against function in a systematic fashion and derive repositioning hypotheses.

4.2.1 Structural descriptor selection

The structure of two chemical compounds can be compared in a variety of ways, depending on the application (Johnson and Maggiora, 1990). In my case, the

main motivation was to find a representative descriptor with a good dynamic range for the dataset considered - a thousand approved drugs. Moreover, the structural descriptor must be relatively easy to handle, fast to compute and with an explicit meaning, solely depending on the molecule and not on external factors.

In order to match these requirements, I decided to focus only on two-dimensional chemical structures, as numerous methods exist to compute them and because they are arguably more accurate for predicting target affinity than three-dimensional descriptors (Nettles et al., 2006), and therefore more suited to study bioactivity. The interpretation of the results is also easier and directly related to the chemical groups present in the structure (Todeschini and Consonni, 2009).

Two-dimensional structures can be represented by fingerprinting methods (see section 4.5.1). Numerous implementations exist, varying in the chemical patterns encoded. I chose to try four of them over my dataset: hybridization, extended, MACCS and PubChem fingerprints (see section 4.5.1). These methods were selected because they are readily available in the Chemistry Development Kit (Steinbeck et al., 2003) and relatively different one from another, therefore providing independent yet comparable results. Figure 4.1 illustrates this: the plot shows the density distribution of similarity values between pairs of drugs for the various fingerprinting methods considered (see section 4.5.2 and 4.5.3).

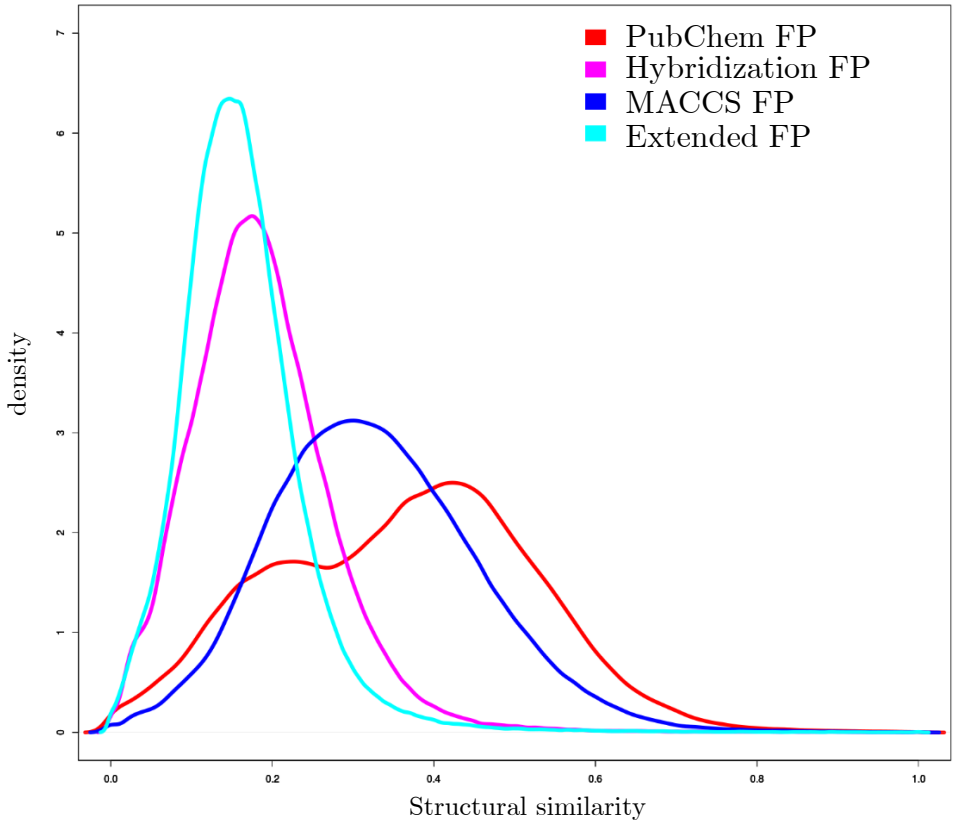


Figure 4.1: Kernel density distribution for various chemical fingerprinting (FP) methods. All the methods have been applied as implemented in the CDK (cf section 4.5.3). Each approved drug was compared against all other approved drugs (pairwise comparison), in order to determine which methodology provides the most suitable distribution to study the dataset.

Different methods have different curves: MACCS and PubChem fingerprinting functions have wider distributions, as shown by the value of the interquartile range, higher than with the other methods (Table 4.1). The distribution of structural similarity values is an important criterion for this analysis, as it reflects the spread of the data. In my case, the wider, the more dynamic and therefore the better.

	PubChem	Hybrid.	MACCS	Ext.	Mean corr.	Qu.1	Mean	Qu.3	Range
PubChem	X	0.65	0.61	0.73	0.66	0.24	0.36	0.47	0.24
Hybrid.	0.65	X	0.66	0.84	0.72	0.13	0.19	0.24	0.11
MACCS	0.61	0.66	X	0.65	0.64	0.24	0.33	0.41	0.17
Ext.	0.73	0.84	0.65	X	0.74	0.11	0.16	0.2	0.09

Table 4.1: Pearson's Correlation values between various fingerprinting methodologies. "Mean corr." stands for the mean value of the correlation coefficients. "Ext." stand for Extended fingerprinter, "Hybrid." for Hybridization. "Range" describes the interquartile range (Qu.3 - Qu.1), based on the similarity values.

Nonetheless, the agreement between the various fingerprinting methodologies also has to be considered: the goal is to find an average structural descriptor, somehow representative and not too polarised, in order to derive systematic conclusions later on. The agreement between methodologies was defined by the Pearson correlation coefficient (see section 4.5.3) and is presented on Table 4.1. Briefly, this coefficient ranges between -1 to 1 and reflects how correlated two series of points are, 1 being a total positive correlation and -1 a total negative correlation. As each method is compared against all other fingerprinting methodologies, I considered the average of Pearson's coefficients as a representative metric; the higher the value, the more a method agrees with the others ("Mean corr." on Table 4.1). From this heuristic, table 4.1 shows that the extended fingerprinting method is the most in line with the others (mean = 0.74), MACCS being the one agreeing the least (mean = 0.64). Nonetheless, all methods have pretty similar average values (see Table 4.1), meaning that all techniques reach overall the same level of agreement. Note that the extended and hybridisation fingerprints have the highest agreement value between them (0.84), reflecting the closeness in the implementation (personal discussion with CDK developers).

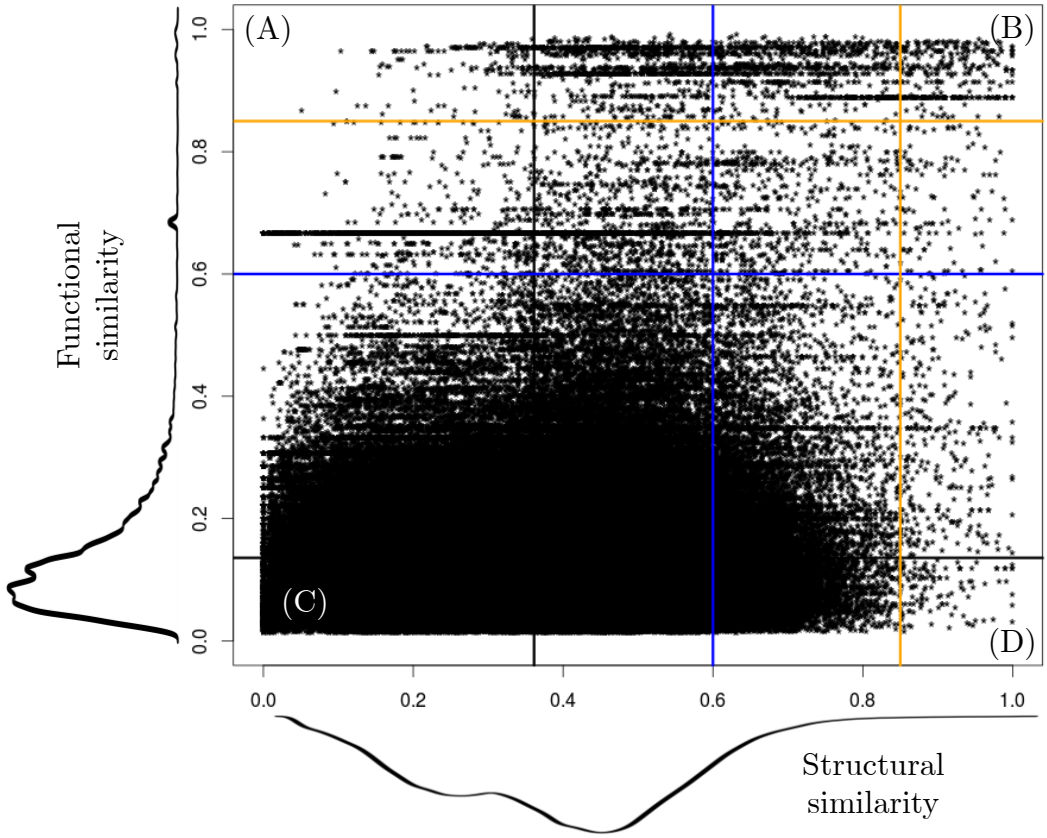
Based on these results, I decided to use the PubChem fingerprint to represent the chemical structure of drugs. The method distributes best the dataset analysed and agrees well with the other fingerprinting methodologies tested, features required for the subsequent drug repositioning analysis.

4.2.2 Dissimilar structures have dissimilar functions

The functional descriptor derives from the structure of the FTC and the semantic similarity. Given a pair of drugs, the closer they are present in the taxonomic tree,

the more similar they are inferred to be (cf section 2.4.3). From this selection of the functional and structural descriptors, it is possible to study the relationship between drugs.

The similar property principle states that similar structures have similar biological activities. The rule was derived from QSAR analyses, where the goal is to try to fit a chemical structure inside a cavity, for instance the active site of an enzyme (Todeschini and Consonni, 2009). In such a case, the rule is intuitively acceptable, yet numerous exceptions are known. The functional descriptor introduced can abstract away from this physical viewpoint and appreciate the similarity relationship in a systematic fashion. In this regard, Figure 4.2 illustrates the distribution of similarity values for all pairs of drugs. Note that the indication is not taken into consideration at this stage.



1.0% 4,457	0.45% 2,060	0.48% 2,222	0.08% 355
93.2% 427,155	5.4% 24,731	99.3% 455,278	0.12% 548

Figure 4.2: Functional against structural similarity values for approved drugs. Each drug is compared against all other drugs (pairwise comparison) using both the structural and functional descriptors and corresponding to one dot or data point on the graph. Two different arbitrary thresholds are applied, represented by the blue and orange lines on the graph. The blue line separates the fairly similar values (>0.6) from the rest, and the orange ones split up the highly similar (>0.85) from the rest of the dataset. The graph is divided and labelled into 4 sections, identified by letters on the figure. The numbers of data points present in each one of these areas are listed on the table below the plot. The kernel density distribution are plotted on the side of the axis (qualitative) in order to appreciate the distribution of the data.

The scatter plot is further divided into areas, broadly separating groups of drugs based on their relative similarity. I chose to consider two different thresholds, in order to separate the similar compounds from the dissimilar ones, repre-

sented by the blue and orange lines in Figure 4.2. The first threshold (blue) is set at an arbitrary similarity value of 0.6. This number appears able to separate the relatively similar from dissimilar compounds, and is derived from observations made in Figure 4-9 (discussed later). The second threshold (orange line) was set at the arbitrary value of 0.85; it separates strongly similar compounds from the rest. This value is generally accepted as a cut-off (Wikipedia (2014a) and personal discussions). The two thresholds reveal the same trend of data distribution (tables in Figure 4.2) in their different areas.

From the distribution of values on the graph, it is clear that the large majority of molecules have dissimilar structures and dissimilar functions. This corresponds to area C on the Figure 4.2, containing either 93% or up to 99% of the data points, depending on the threshold considered. This result can be explained as follows: approved drugs cover a wide range of different bioactivities (dissimilar functions), affecting numerous distinct processes involved in diseases. Their pharmacology is mediated via an interaction with different protein targets, therefore different chemical structures are needed. Moreover, for patent concerns, dissimilar structures are usually sought in order to maximise intellectual protection (Barratt and Frail, 2012). This explanation is consistent with the low number of drugs with similar functions and similar structures (area B on the graph), pairs in agreement with the similar property principle.

Interestingly, a number of drugs are not in line with the similarity rule, represented by the data points in areas A and D. Such pairs have either similar functions with low structural resemblance (area A) or high structural similarities with little shared bioactivity (area D). This observation shows the challenge in drug discovery to relate function and structure: similar MoAs can be obtained with different structures (area A), and just because two structures are similar does not imply that they will trigger the same biological effect (area D).

Two conclusions can be derived from this plot. First, the graph reveals that most drugs have dissimilar structures and dissimilar functions, which I interpret as complementary and in line with the similar property principle. The relation between these pairs of compounds is not informative for drug repositioning and they can later be filtered out. Secondly, the plot enables the identification of exceptions to the similarity rule (areas A and D). These pairs of drugs are of particular interest, as they represent an unexpected pharmacology, not in line with the rule and therefore more difficult to identify. Such pairs are present in

limited quantities, yet they intriguingly more numerous than the pairs respecting the principle and will be investigated in more detail.

So far structure has only been compared to function; yet in order to be more meaningful, the indication of the drugs also has to be considered in the analysis. Intuitively, compounds present in the same therapeutic group should share higher similarity values, reflecting that they target related proteins and have more similar functions (section 3.5.2).

4.2.3 The more specific an indication is, the more similar the function and structure are

The ultimate value of a chemical compound arises from its therapeutic usage. In this regard, the legal indication of a drug is the gold-standard of information. This section further analyses the functional and structural descriptors introduced before and compares them against the legal indication and usage, as represented in the Anatomical Therapeutic Chemical Classification System (ATC). The goal of this comparison is to see whether the descriptors can further provide any meaningful information regarding the indication of a drug, which could be used to retrieve repositioning hypotheses.

According to the similar property principle, drugs present in the same therapeutic group should have on average relatively closer structures and functions. In order to validate this assumption, I considered the five hierarchical levels of the ATC as a representation of the specificity of the indication, with the first or top level of the ATC representing generic and broadly defined indications and the fourth level or bottom level characterising very precisely the indication of a compound (see Figure 4.3).

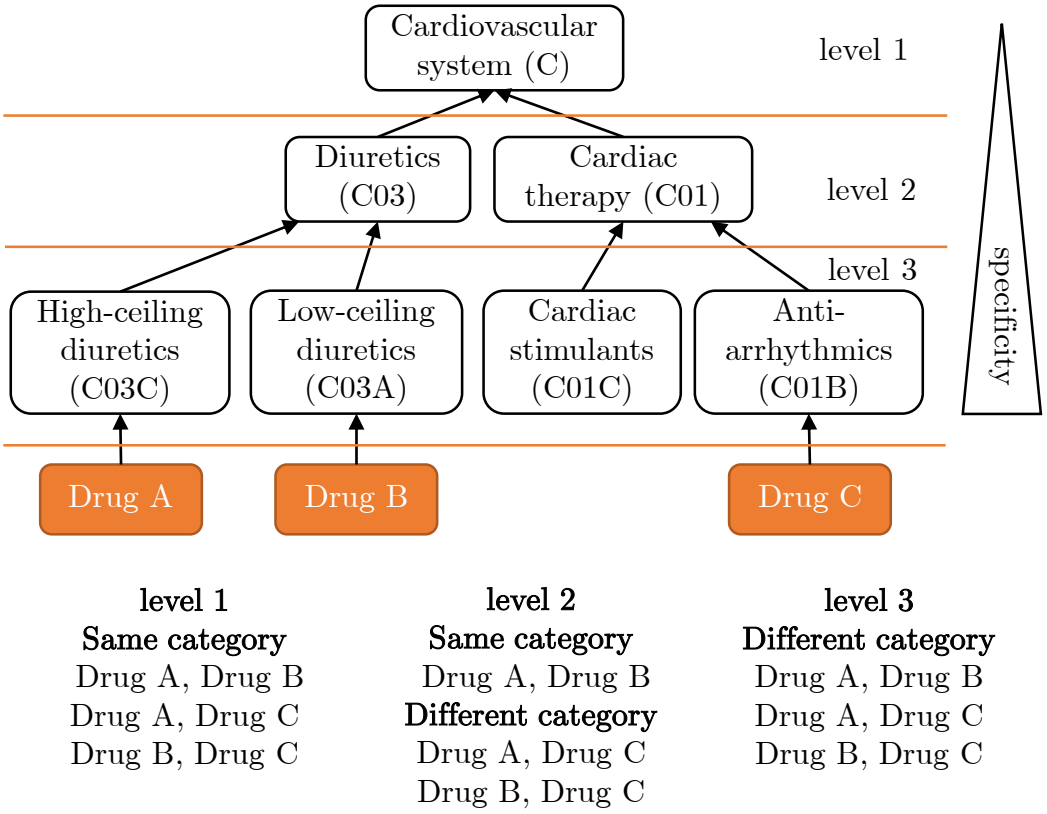


Figure 4.3: Specificity of the indication of a drug as represented in the ATC. The ATC captures drug's indication and is organised over 5 levels (only 4 shown here for clarity), 1 being the highest and most generic level. When descending the tree, the specificity of the indication or action increases. Pairs of drugs can be flagged as belonging to either the same or different ATC category, depending on the level considered (resolution). Three examples are given on the figure in this regards, for the drugs A, B and C. For instance, when only the level 1 is considered, all pairs of drugs are present in the same category (*Cardiovascular system*). When the level 2 is considered, the pair of drugs A and B still belong to the same category (*Diuretics*), but these two drugs are not sharing the indication of drug C, *Cardiac therapy*. At a resolution of the level 3, each drug has a separate indication/action.

It is possible to use this definition of the specificity of an indication to filter pairs of drugs and look at the overall evolution of structural and functional similarities. Such an analysis is shown in Figure 4-4, where only pairs of drugs sharing a common indication are shown. When the specificity of the indication increases, represented by the increasing number of the ATC level considered, the average functional and structural similarity increases too.

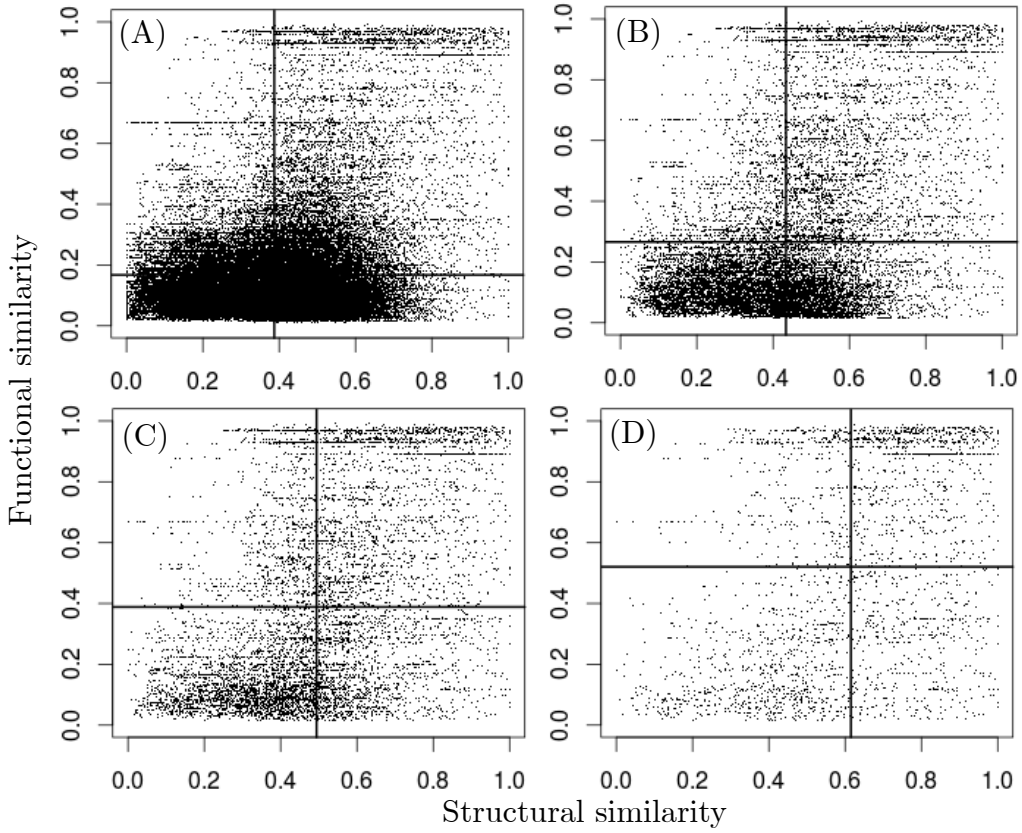


Figure 4.4: Distribution of the functional and structural similarity values for pairs of drugs present in the same ATC categories (same indication). The different panels reflect the increasing specificity of the indication of the drugs. X axes is the structural similarity and Y axes is the functional similarity (calculated as with previous graphs). (A) 1 ATC level resolution. (B) 2 ATC level resolution. (C) 3 ATC level resolution. (D) 4 ATC level resolution. Conceptual explanation of resolution and levels is available on Figure 4.3. When the specificity of the indication increases (resolution increasing), the average functional and structural similarity values increases too (black lines).

On the contrary, when only pairs of drugs indicated for increasingly different indications are compared, the average similarity stays the same, as shown on Figure 4.5 and summarised Table 4.2.

Figures 4-6 and 4-7 respectively show the kernel densities of the structural and functional similarities values for the pairs of drugs sharing an indication at a given ATC level. Note that the two descriptors have different behaviours. The structural similarity appears centred in the middle of the graph, and slowly evolves towards higher similarity values with indication specificity. A jump is observed from level 3 to 4 (brown curve on Figure 4.6), and is manifested by a

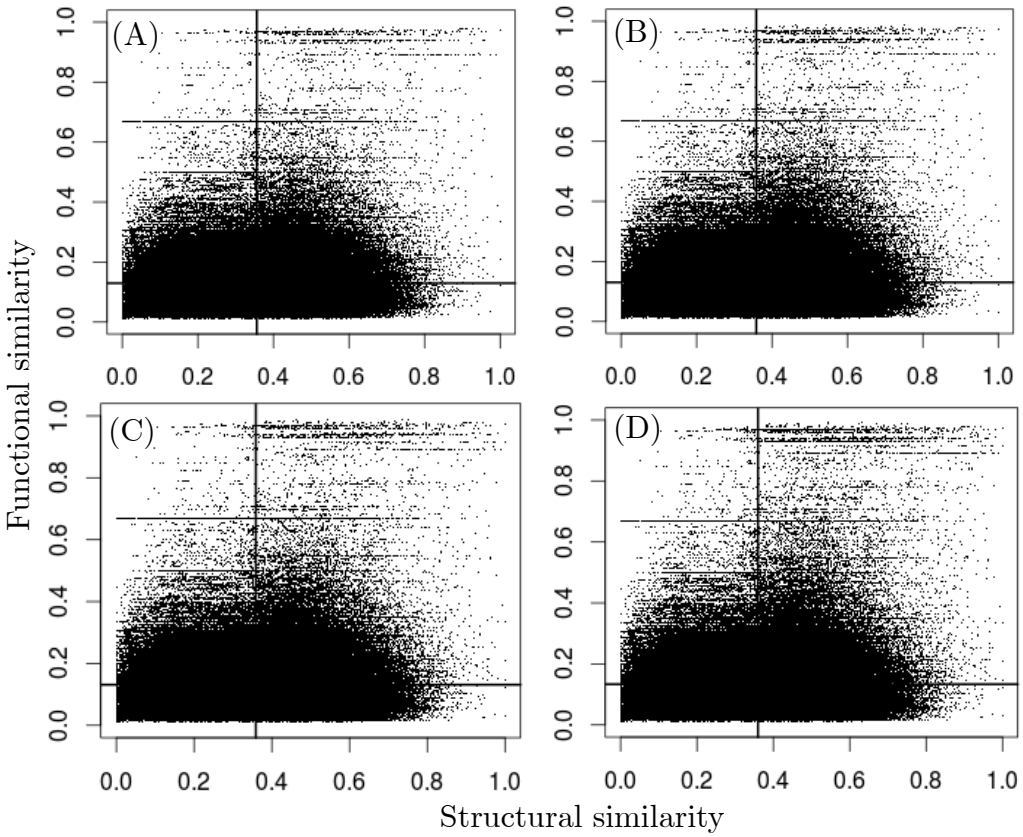


Figure 4.5: Distribution of the functional and structural similarity values for pairs of drugs present in different ATC categories (different indication). The different panels reflect the increasing specificity of the indication of the drugs. X axes is the structural similarity and Y axes is the functional similarity (calculated as with previous graphs). (A) 1 ATC level resolution. (B) 2 ATC level resolution. (C) 3 ATC level resolution. (D) 4 ATC level resolution. Conceptual explanation of resolution and levels is available on Figure 4-3. When the specificity of the indication increases (resolution increasing), the average functional and structural similarity values stays identical (black lines).

larger increase in the average value. This observation is explained by the very definition of the indication coming from the ATC: level 4 handles the categorisation of drugs based on their chemical structures (cf section 3.2.6), therefore it is more likely for a pair of molecules present in the same fourth level ATC category to have very similar structures.

	Specificity of indication (ATC level):	1	2	3	4
Same indication	Average structure similarity	0.39	0.43	0.49	0.62
	Average function similarity	0.17	0.27	0.39	0.52
Different indication	Average structure similarity	0.36	0.36	0.36	0.36
	Average function similarity	0.13	0.13	0.13	0.13

Table 4.2: Evolution of the specificity of indication with the ATC levels. Increasingly similar indications have increasingly similar functional and structural values. The functional and structural similarity values are not evolving when increasingly different indications are considered.

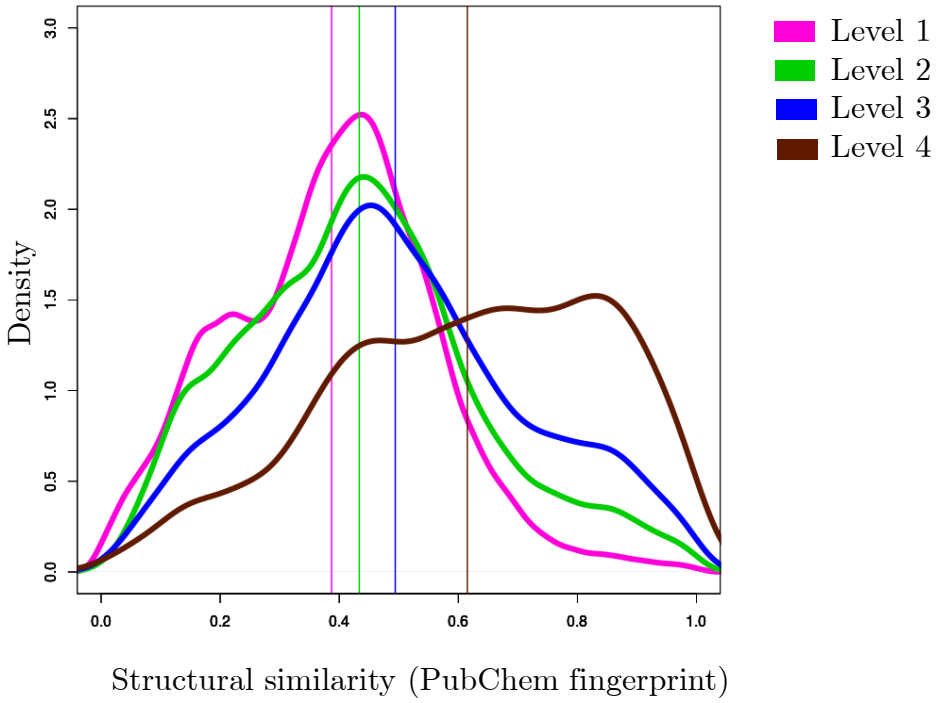


Figure 4.6: Kernel density distribution of the structural similarity values for drugs sharing an indication. All ATC categories have been considered. Each curve represent an ATC resolution level, as indicated on the legend. Conceptual explanation of resolution levels is available on Figure 4.3. Solid vertical lines are the corresponding means. This graph shows that with an increasingly specific indication (increasing resolution) the average structural similarity values increase too.

The functional similarities steadily increase on average (see Table 4.2). With this descriptor, the relative changes in similarities between different ATC levels are more located on the extremes, as shown in figure 4.7: when the specificity

of the indication increases, the number of low similarity values decreases, giving relatively more weight to the high ones.

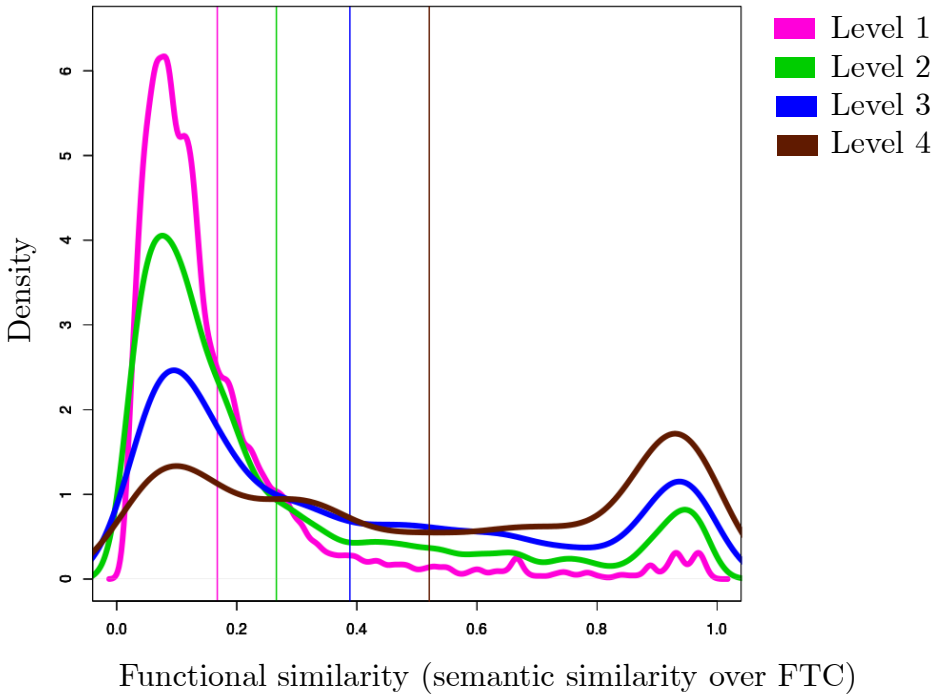
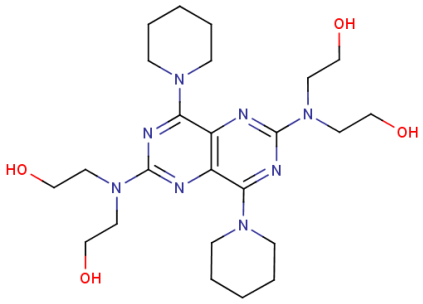


Figure 4.7: Kernel density distribution of the functional similarity values for drugs sharing an indication. All ATC categories have been considered. Each curve represent an ATC resolution level, as indicated on the legend. Conceptual explanation of resolution levels is available on Figure 4.3. Solid vertical lines are the corresponding means. This graph shows that with an increasingly specific indication (increasing resolution) the average functional similarity values increase too. See Table 4.2 for values.

Taken together, these results confirm the similarity principle at a systematic scale: on average, drugs with closely related indications are structurally and functionally similar, as captured by the descriptors used. As an example, the plot showed in Figure 4-4 panel D displays the highest average similarity, when the indication is the most specific (highest ATC level). This average similarity value was used to set the threshold level at 0.6 in Figure 4.2; it separates functionally similar compounds from the rest. Interestingly, on the same graph (Figure 4.4-D), some of the pairs still have low similarity values (structural and functional lower than 0.3); these points are outliers to the similar property principle and reflect the limits of the descriptors. I performed an error analysis in order to

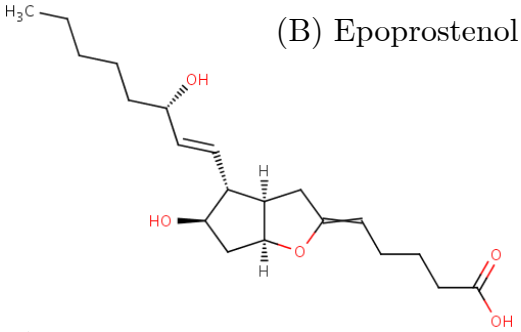
identify the reasons behind the non-respect of the rule by these drugs. The pair of drugs dipyridamole and epoprostenol is a good illustration of the most common cases of misclassification. These two drugs have a relative functional similarity of 0.10 and a structural similarity of 0.14, despite being both categorised as *Platelet aggregation inhibitors* in the ATC (code: B01AC). Although resulting in the same biological outcome and clinical usage, these two drugs are mainly targeting different proteins, a phosphodiesterase in the case of dipyridamole and the P2Y purinoceptor 12 in the case of epoprostenol. In order to interact with these receptors, two different chemical structures are needed and it is therefore expected that the two molecules would have different structures (see Figure 4-8). However, the MoAs from the FTC are not shared either, which comes from missing annotations on the protein targets or because the molecular root of the effect is unknown. It is therefore also not possible to relate these two drugs based on their functions, because of lack of recorded knowledge.

(A) Dipyridamole



- Anti-adenosine deaminase activity agent
- Anti-3,5-cyclic-AMP phosphodiesterase activity agent
- Anti-cGMP-stimulated cyclic-nucleotide phosphodiesterase activity agent
- Anti-protein binding agent
- Anti-zinc ion binding agent
- Anti-cAMP binding agent
- Anti-cGMP binding agent
- Pro-adenosine receptor signaling pathway agent

(B) Epoprostenol



- Anti-nitric oxide biosynthetic process agent
- Anti-inflammatory response agent
- Anti-platelet-derived growth factor receptor signaling pathway agent
- Anti-smooth muscle cell proliferation agent
- Pro-angiogenesis agent
- Pro-guanyl-nucleotide exchange factor activity agent
- Pro-protein binding agent
- Pro-cAMP biosynthetic process agent
- Pro-prostaglandin-I synthase activity agent
- Pro-cAMP-mediated signaling agent
- Pro-heme binding agent
- Pro-peroxisome proliferator activated receptor signaling pathway agent
- Pro-execution phase of apoptosis agent

Figure 4.8: Example of pair of drugs with low structural and functional similarity values, yet classified in the same ATC category and used for the same clinical indication (*platelet aggregation inhibitors*). (A) Chemical structure and list of FTC categories inside which the dipyridamole was classified. (B) Molecular structure and list of FTC categories inside which the epoprostenol molecule was classified. These two drugs do not share any of the FTC categories listed (functional similarity = 0.10) and their molecular structures are dissimilar (structural similarity = 0.14).

The figures presented in this section show an observation of primary importance for drug repositioning: the computational descriptors used are able to serve as a proxy for the expected behaviour of the concepts of function and structure in regards to the real clinical indication of a drug. The more similar a pair of drugs are based on either their functional or structural features, the more likely these drugs are clinically indicated for the same sets of diseases.

Note that the data analysis shown in the two previous sections only identified average patterns; particular data points were not taken into considerations. Some exceptions or outliers exists, which will be considered as starting points to formulate drug repositioning hypotheses.

4.2.4 Isolation of drug repositioning hypotheses

As motivated before (section 3.2.6), the ATC can be considered as a gold standard resource for representing the therapeutic area of a drug. The second level of this classification system in particular represents the clinical indication of a compound well (Organization et al., 2006); it is not too specific to capture the MoA and exact pharmacology, yet still biologically meaningful and not too abstract. Example of second level ATC categories are: *Anabolic agents for systemic use* (A14) and *Anti-haemorrhagics* (B02). I and others (Campillos et al., 2008) (Napolitano et al., 2013) have considered that drugs categorised under different second-level ATC codes can be considered as indicated for different diseases.

Drugs used in the treatment of different conditions are expected to have dissimilar functions, as in principle such agents affect separate biological processes, themselves related to distinct diseases. The graph in Figure 4.9-A recounts this statement: as expected most of the pairs have low structural and functional similarity values.

However, some data points on this plot have relatively high functional similarity values (similarity superior to 0.6, identified by the blue line). These pairs of drugs are unexpected; such compounds appear to be strongly related on the functional level, yet in practice they are clinically used for radically different indications. Such data points have been interpreted as repositioning hypotheses.

In this regard, the values shown above the blue line (0.6) in Figure 4.9-A have been first isolated (see Figure 4.9-B). Some drugs with little knowledge have been discarded (low number of annotations - see section 4.5.6); the remaining set is displayed in panel (C). The pairs of molecules in the area Figure 4.9-C, defined as drug repositioning hypotheses, were isolated from the rest of the data points and further displayed through a web application (`https://www.ebi.ac.uk/chembl/research/ftc-hypotheses/`) in order to facilitate their exploration and validation.

A total of 797 pairs of points or repositioning hypotheses were considered. The hypotheses are grouped by therapeutic areas in the web application. For example, the ATC category J01 - *Antibacterials for systemic use* contains 72 associations or potential drugs that are not indexed as antibacterial and yet could be used as such, based on their known role in the human body (`https://www.ebi.ac.uk/chembl/research/ftc-hypotheses/code/J01`). The web application further provides

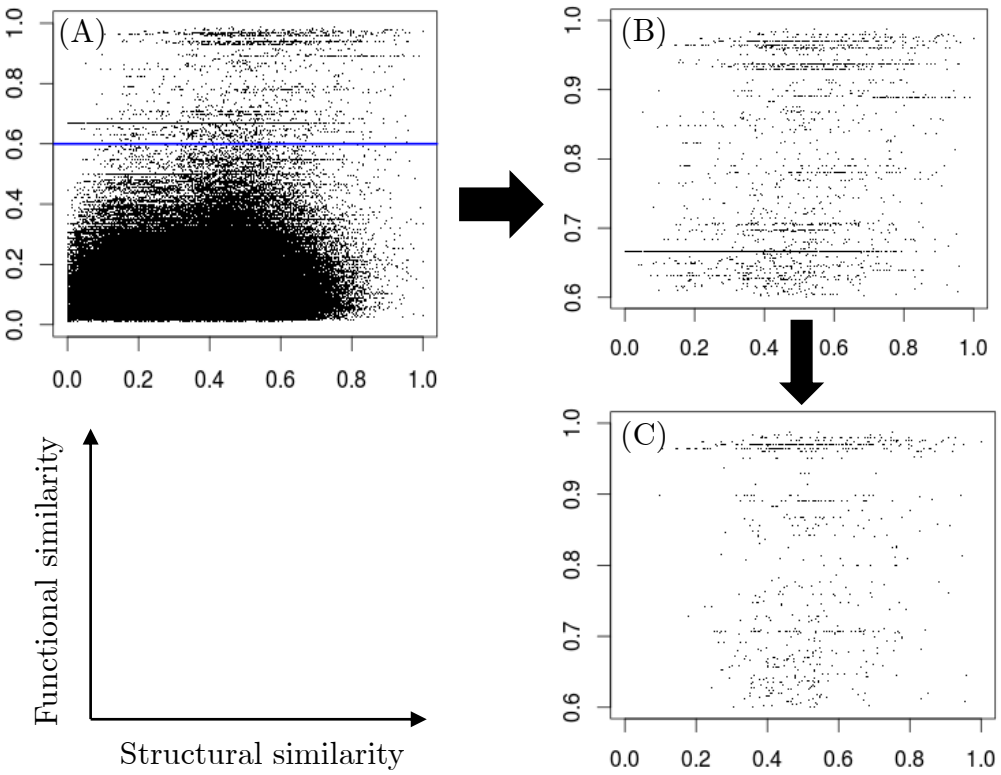


Figure 4.9: Isolation of drug repositioning hypotheses. (A) All the pairs of drugs classified in different ATC categories (level 2 resolution) are plotted (identical plot as Figure 4-5 panel B). The blue line is the threshold above which pairs are functionally similar but are indicated for different usage according to the ATC. (B) Zoom on the data set above the blue line. (C) Pairs of drugs with little recorded knowledge are discarded from the analysis (see material and method). This set of data represents the repositioning hypotheses and are featured in the web application.

an aggregation of the ATC categories related to the one currently investigated. For instance the hypotheses related to the category *Antibacterials for systemic use* are from drugs present in the ATC group L01 - *Antineoplastic agents* (54 hypotheses), S02 (*otologicals*, 16 hypotheses), V03 (*All other therapeutic products*, 9 associations), D06 (*Antibiotics and chemotherapeutics for dermatological use*, 9 hypotheses) and S03 (*Ophthalmological and otological preparations*, 8 hypotheses).

Very often, the same drug is involved in multiple hypotheses. For instance, the antibiotic pefloxacin is functionally similar to 9 other drugs; this results in a network of interconnected drugs related to one therapeutic category. From the web application, it is possible to export the file containing hypotheses to further analyse the graph in Cytoscape (Shannon et al., 2003), a program commonly used

to visualise networks and systems. These graphs will be extensively discussed in the rest of this chapter.

From the relationship between the structure, function and indication of approved drugs, it is possible to isolate outliers not following the predicted behaviour. Such data points extracted from the analysis have been defined as drug repositioning opportunities and are available via a web interface. Next sections will focus on particular therapeutic groups and use-cases, in order to identify the relevance and the limits of the approach.

4.3 Open drug repositioning hypotheses

The ATC provides data regarding the anatomical group inside which a drug has been assigned (first level) as well as its main therapeutic area (level 2). I considered this information as a gold standard defining the indication of a drug. As a reflection of polypharmacology, some of the drugs are already present in multiple ATC categories, or have multiple indications. For instance, acetylsalicylic acid is classified as a stomatological preparation (A01), an antithrombotic agent (B01) and an analgesic (N02), according to the various roles it can play in the human body. The hypotheses generated render a similar result: drugs are predicted to be active in other ATC categories on top of the ones they have already been manually assigned to.

The aim of this section is to characterise and interpret the hypotheses; indeed, before moving forward to perform laboratory experiments, it is necessary to have a clearer idea of the molecular reasons behind the predictions.

First I will present an overview of the hypotheses and illustrate how drug repositioning relates to the non-regulated uses of drugs. This topic is known as off-label uses (Wikipedia, 2014c) and concerns around 20% of the drug's prescriptions (Stafford, 2008). An off-label use is the indication of a drug for a non-regulated usage, either in terms of therapeutic group, dosage or patient group. This practice is legal in most countries, yet as it is non-regulated there is little information about it.

Secondly, I will investigate drug repositioning hypotheses for cardiovascular hypertension and Alzheimer's disease. These use-cases are also an opportunity to compare the hypotheses generated using functional similarity (section 4.2.4)

against the *toolbox* approach, which considers drugs classified inside FTC categories directly as relevant for a biological process.

4.3.1 Relationship between therapeutic areas

In the context of this work, a repositioning hypothesis is defined as *a pair of drugs that are functionally similar yet currently indicated for different clinical purposes*. In practice, this relationship can serve to relate therapeutic groups: some hypotheses are present more often than others between two given groups, providing insights as to how close these clinical areas are. In this respect, Figure 4-10 shows the association between therapeutic areas, with a resolution of two ATC levels.

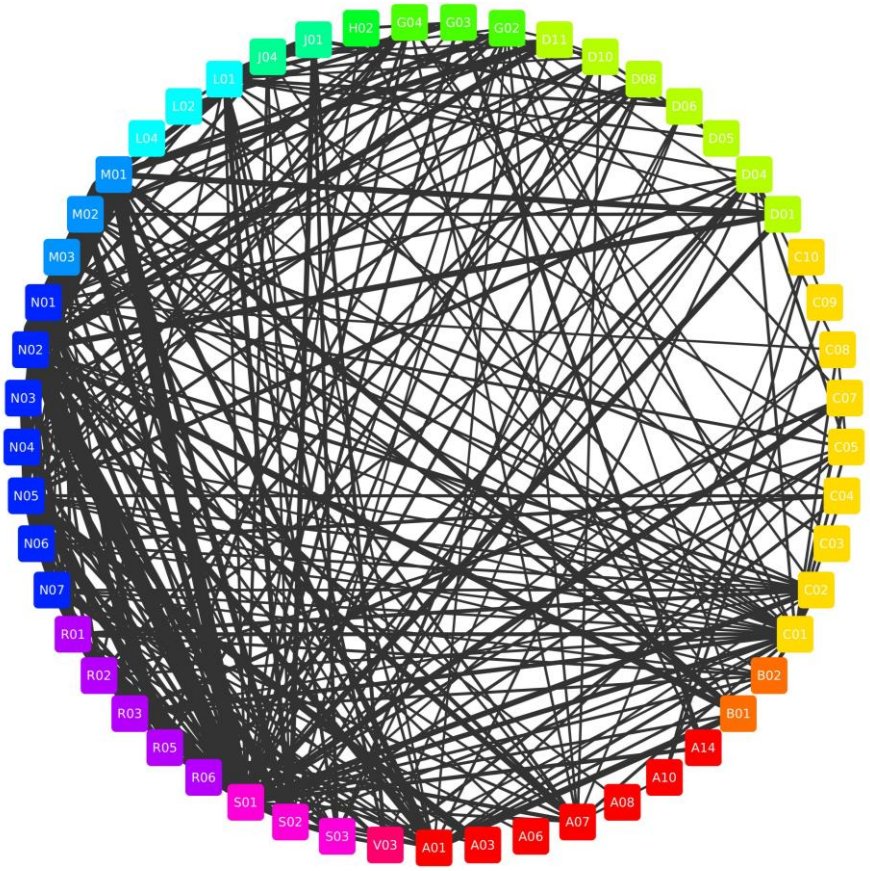


Figure 4.10: Raw associations between therapeutic areas. All the repositioning hypotheses between two a therapeutic groups are merged to define the strength of association amongst these groups. The boldness of the line is directly proportional to the strength of association. For instance, there exists more repositioning hypotheses between the groups M01 and S01 (very bold line) than between the groups C10 and A07 (thinner line). Categories are sorted on alphabetical order and the colour is assigned based on the first letter of the group (therapeutic area). All the group S is pink for instance (S01, S02 and S03).

It is rather challenging to derive any meaningful conclusion from the graph, yet the hairball reveals an interesting features of the dataset: a high inter-group connectivity, which appears stronger in some cases and weaker in others. For instance, D categories (dermatologicals) are less well connected than N series (nervous system), simply from visual observation. In order to simplify the problem faced and extract the essence of the dataset, I further decided to group categories considering solely their first ATC level and filtering out weak associations (cf methods section 4.5.7). The result is shown in Figure 4.11.

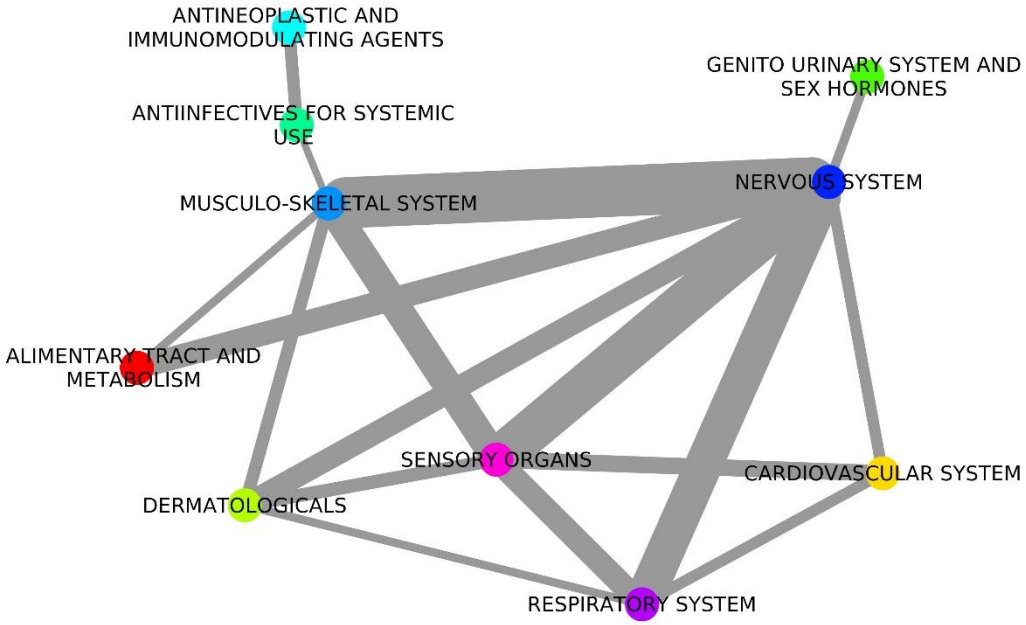


Figure 4.11: Filtered and abstracted associations between therapeutic areas. The network is built from the data point shown on Figure 4-10. ATC categories have been grouped based on their first level: for instance all the hypotheses for the groups S01, S02 and S03 have been merged as S (or *sensory organ* in this figure). Moreover, only the links involving a minimum of 25 hypotheses (arbitrary) have been kept in order to filter out the weak associations. The network shows the areas subject to drug repositioning or for which there exists a "promiscuous" pharmacology, which can be captured by the degree of a node (number of connections with other nodes).

The resulting graph abstracts away details and shows clearer high level associations between groups. The most connected node is the nervous system (degree of 6). The high connectivity around this therapeutic area could be interpreted as the presence of numerous repositioning hypotheses; however, it is much more likely to be due to off-label usages. Neurology is a domain particularly famous for this (Cras et al., 2007). For instance, it is genuinely difficult to define and quantify concepts such as *pain* (Bonica et al., 1979) or any mental illness, therefore clinicians do not always use the officially indicated drug, depending on their diagnostic, personal experience and history of the patients (personal discussion with clinical pharmacologists).

Other therapeutic areas such as the ones related to sensory organs, dermatology and the musculoskeletal system are also well connected (degree of 4 and 5). The presence of the musculoskeletal category can be explained by the promiscu-

ity of anti-inflammatory agents; usually such drugs are also active on the nervous system, explaining the strong relation between these two groups. This high level overview can be refined with the help of Figure 4.12.

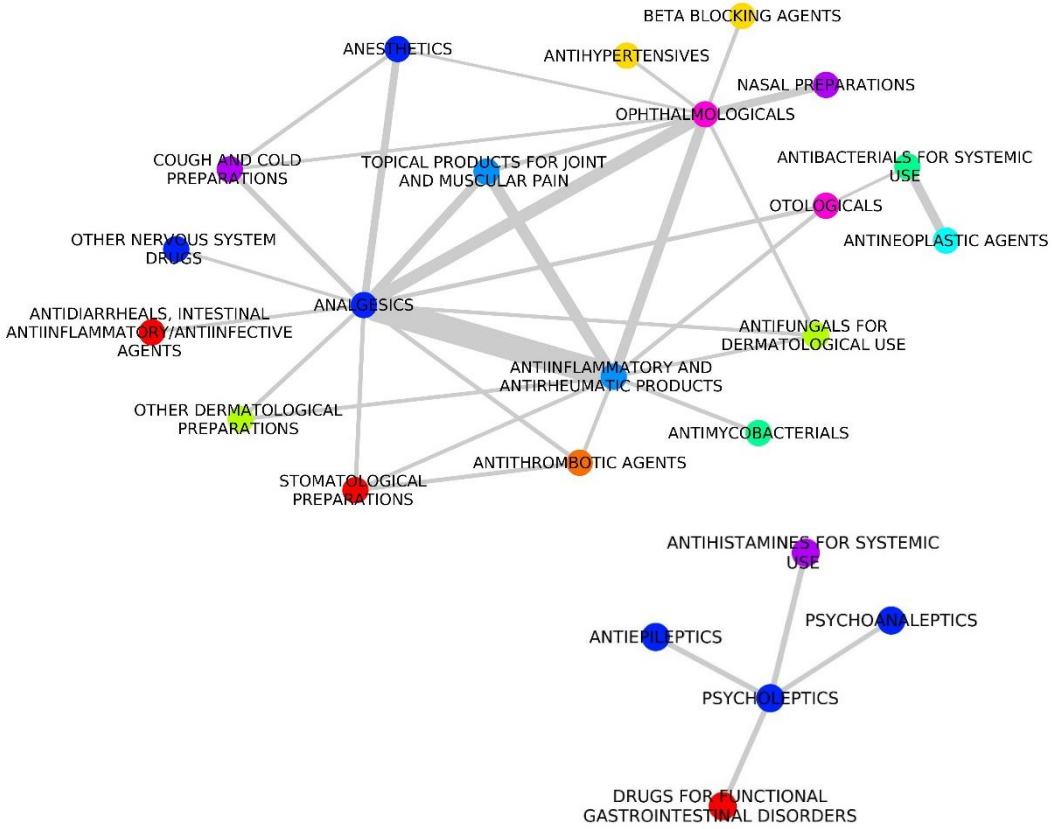


Figure 4.12: Filtered associations between therapeutic areas. The network is built from the data point shown on Figure 4.10. Weak associations (less than 15 hypotheses) have been filtered out. The graph shows the high level association between therapeutic areas, for a two ATC level resolution. See text for interpretation of the edges.

These graphs focus on the ATC level 2 indications, as in Figure 4.10, but with weaker associations filtered out (section 4.5.7). This more granular view of the data confirms the fussiness around neurological compounds; 6 out of 26 categories are from this therapeutic area and are well connected to the other nodes. Some associations are of special interest, as summarised in Table 4.3 .

Analgesics (relief from pain) are strongly associated with cough and cold preparations. This pair is not so surprising; anti-flu medicines often mediate their effect as palliative pain killers, as confirmed in the ATC description of the category (ATC, 2014b). The strong connection between anti-inflammatory drugs

Therapeutic groups association		Interpretation
Cough and cold	Analgesics	Analgesics are used in cough preparations (ATC, 2014b)
Anti-inflammatory	Antimycobacterials	No naive interpretation
Analgesics	Ophtalmological	Related MoA, regulation of blood pressure
Ophtalmological	Nasal preparation	
Antihypertensive	Ophtalmological	
Beta blocking agent	Ophtalmological	
Antibacterial	Antineoplastic	Ortholog targets, replication of DNA
Antihistamines	Psycholeptics	Drowsiness, specially with first and second generation (Gengo and Gabos, 1987)
Gastrointestinal disorder	Psycholeptics	Currently used as treatment, preparations contain psycholeptics (ATC, 2014a)
Analgesics	Anti-inflammatory	Side-effect, shared bioactivity (Hunskaar and Hole, 1987)

Table 4.3: High level association between therapeutic groups and shared pharmacology. See Figure 4.10 for source network.

and analgesics observed in the Figure 4-10 is confirmed; compounds from either of these classes are known to have overlapping pharmacology (Hunskaar and Hole, 1987). This relationship is further supported by the extra links to *topical products for joint and muscular pain*, showing the ambiguity of biological action and the relatedness of these indications.

Another neurological category, psycholeptics (calming effect), is related to antihistamines and drugs for gastrointestinal diseases. The former association explains the drowsiness side-effect provoked by first and second antihistaminic agents. For gastrointestinal disorders, the link is explainable by the fact that psycholeptics are often administered as a palliative solution for such dysfunctions (ATC, 2014a). An edge also connects stomatological preparations to analgesics, supporting this observation.

The strong relation between anti-neoplastic agents and anti-bacterials can be explained by the orthology between the proteins targeted by these drugs, namely the DNA replication machinery. The replication apparatus of bacteria and humans are still fairly similar (Hurle et al., 2013), therefore a drug inhibiting one can also inhibit the other.

Ophthalmologicals are connected to cardiovascular drugs (beta blocking and antihypertensives agents) as well as to nasal preparations. This hub originates because of the shared fundamental MoA between all these categories: regulation of blood pressure and flow, in different systems. The functional similarity between these categories is high, as the pharmacological effect is mediated via the same

biological process, yet the administration and delivery of the drug in the body leads to the specificity of the indication. This relationship will be analysed in more details in coming sections.

Finally, some of the edges are difficult to interpret without looking at the underlying pairs of drugs. For instance, the association between anti-inflammatory and anti-mycobacterial agents is unusual and requires deeper investigation to be interpreted (not discussed in this document).

This overview of the relationship between therapeutic groups reveals some expected associations, confirming the validity of the approach. Some known cases are represented in the graphs shown and are documented in the scientific literature. Polypharmacology between groups appears non-homogeneous, as some therapeutic areas are closer than others in terms of functional similarity and strength of association. The list of drug repositioning hypotheses contains some off-label pairs, which are drugs known and already used for an alternative indication yet not formally approved for it.

From this generic presentation of the dataset and in order to better characterise the repositioning hypotheses, I decided to focus in more detail on agents active for treatment or prevention of cardiovascular hypertension. Finally I demonstrate how the categories of the FTC can be used as a starting point to address Alzheimer's disease (toolbox approach).

4.3.2 Drug repositioning for hypertension and the cardiovascular system

The beginning of this chapter (section 4.2.3) compared pairs of drugs considering the relative similarity of their functions and structures. The closeness between two roles was defined as the semantic similarity over the content of the FTC; such a metric helped to identify related compounds, forward repositioning hypotheses and look at the relationships between therapeutic groups. This approach, referred to as *similarity-based*, is useful when a known therapeutic category already exists in the gold standard (ATC) which can be used as a starting point for further investigations.

Complementarity, the FTC can also be directly used as a toolbox, in order to find drugs to fix or repair a molecular machinery; some FTC categories contain active molecules in theory capable of producing the described pharmacological

effect. Therefore, knowing the desired MoA can help to identify what molecule to use for a particular condition. This methodology is named the *toolbox approach*. The difference between the two approaches are illustrated and summarised in Figure 4.13.

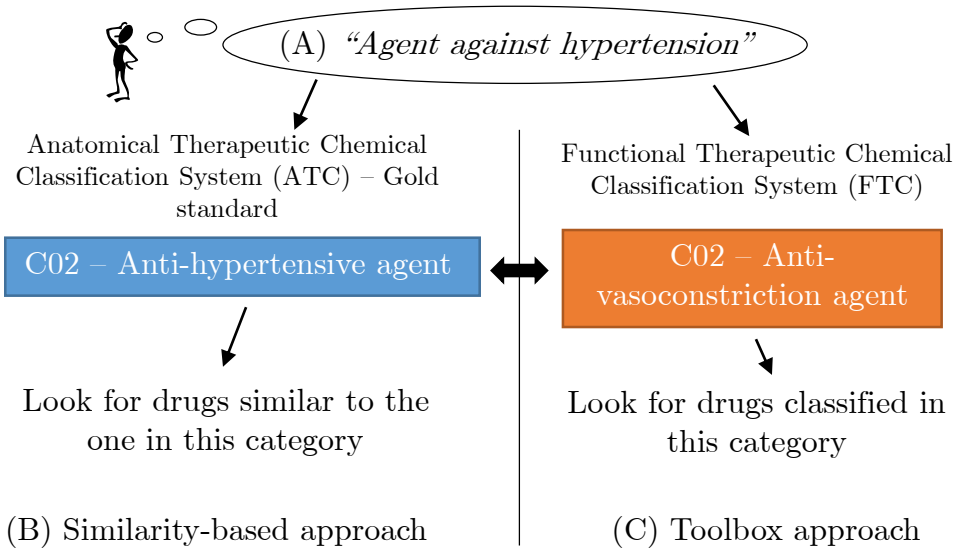


Figure 4.13: Comparison of the similarity-based against toolbox approach for drug repositioning hypotheses investigation. (A) The concept formulated in natural language as in the head of a researcher or mentioned in a text is formalised and mapped to a term from a classification, either the ATC or the FTC in this case. (B) Similarity-based approach: the drugs functionally similar (semantic similarity over the FTC) to the ones listed in the ATC category are the drug repositioning hypotheses (displayed in the web application). (C) Toolbox approach: the drugs classified under the selected FTC category are capable of producing the MoA of interest. These compounds have been automatically assigned to the categories during the reasoning step (see Chapter 3 section 3.2.5). Note that often the ATC and FTC category have equivalent meanings, as emphasized by the bold black arrow between the categories.

This section demonstrates how the FTC and the previously generated drug repositioning hypotheses can be used in practice, using hypertension as an example.

High blood pressure, or hypertension, is a medical condition leading to an increased risk of strokes, heart attacks and cardiovascular diseases (Law et al., 2003). In Europe alone, this condition is considered epidemic; hypertension affects between 30 to 45% of the whole population (Swedberg et al., 2005), making it one of the leading causes of death. From a mechanistic perspective, hypertension can be straightforwardly regulated by lowering the blood pressure in the

affected patient; this biological outcome can however be achieved in a variety of ways. For instance, the physical contraction of cardiac cells can be directly inhibited by blocking a receptor, the signalling pathway triggering the vasoconstriction can be perturbed, or the density of the blood itself can be reduced in order to make it more fluid (Swedberg et al., 2005). All these different mechanisms are of interest because they produce the same therapeutic outcome, with yet different safety profiles or patient groups.

The FTC contains and defines such mechanisms, therefore the resource can assist in choosing the right drug to elicit the desired action (toolbox approach). The commonly considered mechanisms and MoAs for hypertension treatment are presented below, alongside their corresponding categories in the FTC and ATC, when existing. Moreover the repositioning hypotheses generated earlier in this chapter using the similarity-based approach will be discussed when relevant.

The hypertension use-case exemplifies a detailed mechanistic analysis that can be performed over the FTC dataset, from the functional representation to drug repositioning hypotheses.

4.3.2.1 Antihypertensives

A feature of the FTC is its ability to abstract away from specific chemical interactions in order to solely focus on high level biological processes. In this respect, at least two FTC categories - or MoAs - are of interest to address hypertension: *Pro-vasodilation agent* (`https://www.ebi.ac.uk/chembl/ftc/FTC_P0042311`) and *Anti-vasoconstriction agent* (`https://www.ebi.ac.uk/chembl/ftc/FTC_A0042310`). The former type contains only one drug, whereas the latter features 49 compounds. All of these drugs are deemed to be active for the condition studied and some of them are already indicated as such (e.g, prazosin), whereas others are currently used for different purposes, such as clozapine, an antipsychotic agent. The FTC categories can be used to identify new compounds capable of producing the biological effect following this toolbox approach; no previous knowledge is necessary, except about the name of the mechanism studied in order to find active drugs (see Figure 4.14).

The semantically equivalent ATC category is *Antihypertensives* (C02) and contains single molecules and drug combinations. The drug repositioning hypotheses for this ATC class are listed and interpreted in Table 4.4 (`https:`

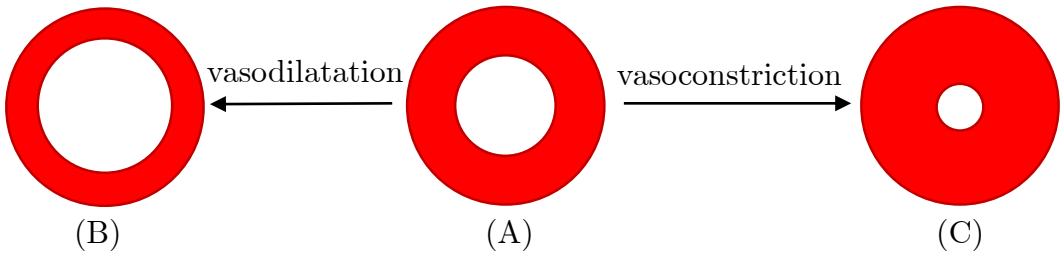


Figure 4.14: Schematic representation of the dilatation/constriction process of blood vessels. (A) Cross section of blood vessel in normal state. (B) Vasodilatation process, the size of the blood vessel increases, reducing the overall blood pressure (less tension). (C) Vasoconstriction process, the diameter of the blood vessel diminishes, increasing the overall blood pressure. Preventing or treating hypertension can be achieved by either promoting the state (B) (Pro-vasodilatation agents) or by preventing state (C) (Anti-vasoconstriction agents).

`//www.ebi.ac.uk/chembl/research/ftc-hypotheses/code/L02`).

The presence of neurological drugs is explainable by their adrenergic side activity; such molecules bind receptors that are similar to the ones involved in blood pressure regulation, yet are located in different parts of the body and produce a different therapeutic effect. A similar pattern is observed with ophthalmological agents, administered for glaucoma, a disease closely related to eye hypertension (Quigley, 1996). In this case too, vasodilators are prescribed, but the delivery and formulation of the molecule confers the specificity of the effect. This evidence supports the validity of the hypotheses; administering the drugs in the cardiovascular system rather than in the eye would probably reduce the overall blood pressure.

A series of arguably good hypotheses are related to compounds used against nasal congestion. Their reported pharmacological effect, vasoconstriction of the nose's arteries (Bende et al., 1996), is the opposite of the one wanted. However, the biological activity of such compounds is strongly related to that of epinephrine, which is particularly sensitive to concentration (Evans et al., 2001). Depending on the dosage and the body part concerned, the epinephrine molecule can produce radically different outcomes, from vasoconstriction to vasodilatation. On this basis, the repositioning hypotheses related to nasal preparations seem more plausible; still, the delivery methods would need to be adequately optimised.

The information in the FTC helps to group drugs by MoAs as expected,

Identifier reference drug	ATC code reference drug	Identifier hypothesis	ATC code hypothesis	Hypothesis interpretation
DB00457 (Prazosin)	C02 (Antihypertensives)	DB01162 (terazosin)	G04 (Urologicals)	Valid hypotheses, the drug is used as such already
DB00590 (doxazosin)	”	”	”	”
Centrally active hypotensive agents (group A)	”	DB00668 (Epinephrine)	A01 (Stomatological preparations) B02 (Antihemorrhagics) C01 (Cardiac therapy) R01 (Nasal preparations) R03 (Drugs for obstructive airway diseases) S01 (Ophthalmologic)	Drug involved in numerous process, effect varies with quantities, to verify pro-drugs and how it influences
”	”	DB00368 (Norepinephrine)	C01 (Cardiac therapy)	Related therapeutic group, central hormone
”	”	DB00697 (tizanidine)	M03 (Muscle relaxants)	Management of spasticity, same receptors but different locations, inhibition of motoreurons
”	”	DB00320 (Dihydroergotamine)	N02 (Analgesics)	Migraine therapy, vasoconstrictor
”	”	DB00413 (Pramipexole)	N04 (Anti-parkinson drugs)	For Parkinson’s disease and restless syndrome. Used off-label for cluster headache. Effect on vaconstriction?
”	”	DB00633 (Dexmedetomidine)	N05 (Psycholeptics)	Because of its pain killing action, it also decreases the blood pressure and heart rate, different location of receptor triggers different effect
”	”	DB01577 (Methamphetamine)	N06 (Psychoanaleptics)	Plausible, illegal molecule
”	”	DB04948 (Lofexidine)	N07 (Other nervous system drugs)	Adrenergic agonist, reported as short-acting antihypertensive, but mostly used to relieve opiate dependency, usage already reported
”	”	DB00935 (Oxymetazoline)	R01 (Nasal preparations) S01 (Opthalmologicals)	Vasoconstrictor to relieve nasal congestion. Also used for ocular inflammation.
”	”	DB06711 (Naphazoline)	”	”
”	”	DB06694 (Xylometazoline)	”	Nasal vasoconstrictor decongestant
”	”	DB00397 (Phenylpropanolamine)	R01 (Nasal preparations)	Produces vasoconstriction
”	”	DB00964 (Apraclonidine)	S01 (Opthalmologicals)	Used in glaucoma therapy, anti-hypertensive drug in the eye system
”	”	DB00484 (Brimonidine)	”	”
”	”	DB00449 (Dipivefrin)	”	”

Table 4.4: Drug repositioning hypotheses related to hypertension (ATC code C02). The symbol ” refers to the value of the previous cell.

without consideration for the anatomical part acted upon. In order to be repurposed, the delivery mechanisms of the drugs need to be adapted to transport the molecule to the right place in the body. The FTC provides abstract level MoAs, such as the ones presented in this section. It is also possible to use the resource to investigate more detailed mechanisms, as discussed in the coming parts.

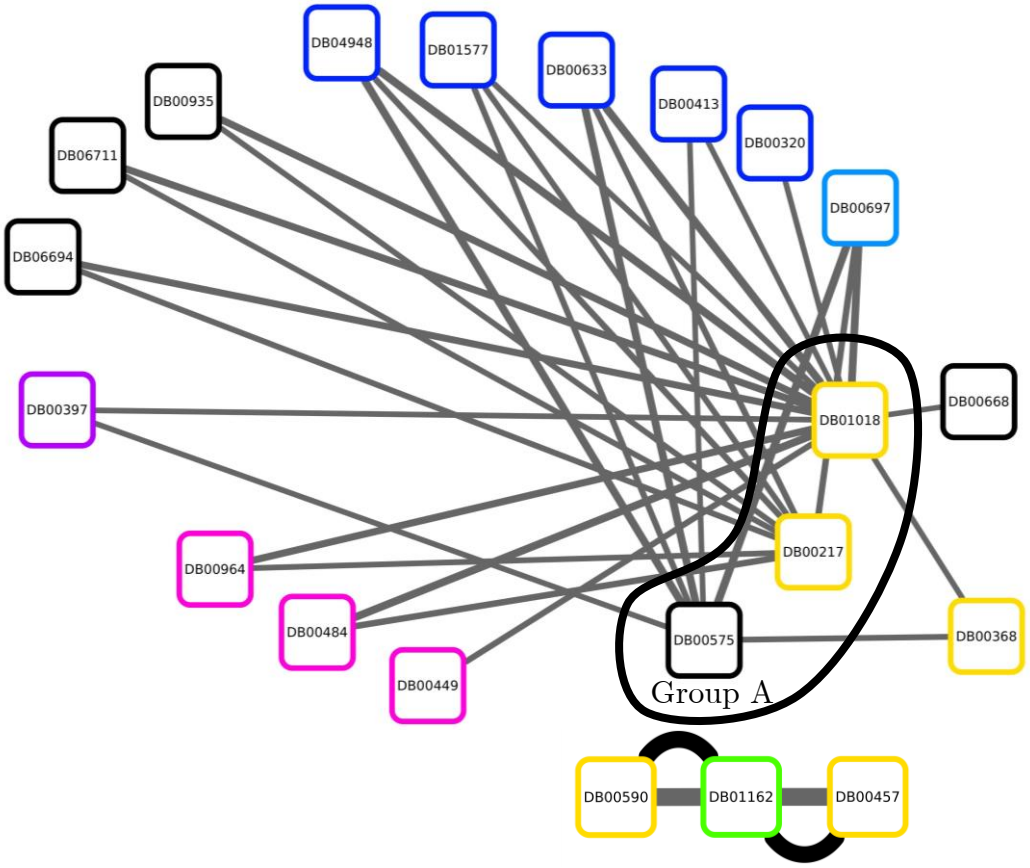


Figure 4.15: Drug repositioning hypotheses related to the ATC category C02 - *Antihypertensives*. The width of grey edges is proportional to the functional similarity values between drugs. Legend for colours and drug name are available online and in Table 4.4.

4.3.2.2 Calcium channel blockers

One discovered way to reduce blood pressure is by blocking the entry of ions (calcium mostly but also potassium and chloride) into cardiac muscle cells (Swedberg et al., 2005). Deprived from the ions, the cardiac rhythm decreases and the arteries dilate, leading to a drop in blood tension (Figure 4.16).

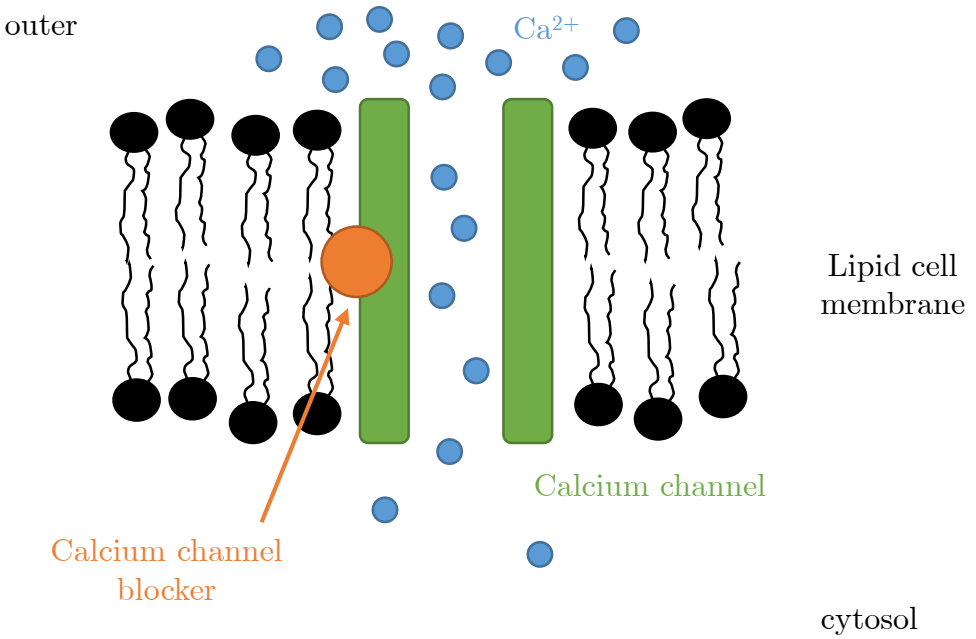


Figure 4.16: Schematic illustration of the mechanism of action of calcium channel blockers. The calcium ion is necessary for the contraction of the heart muscle; its releases from internal cellular store or entry inside the cell triggers the contraction. By blocking the channel, the calcium does not circulate any more and the cardiac rhythm decreases, therefore reducing the blood pressure and improving a hypertension.

One FTC mechanism is directly related to this action: *Anti-high voltage-gated calcium channel activity agent* (FTC_A0008331 - 13 drugs). As expected, most of the drugs present in this category are indicated as antihypertensive agents (12 out of 13). The direct superclass of this category is the more generic concept *Anti-voltage-gated calcium channel activity agent* (FTC_A0005245 - 36 drugs). Interestingly, inside this broader class, a quarter of the drugs (8/36) are anticonvulsants (anti-epilepsy) and further classified as *Anti-low voltage-gated calcium channel activity agent* (FTC_A0008332). This observation reflects the therapeutic closeness between this series of compounds; the two channel types (low and high voltage) can produce radically different physiological outcome from similar structures. The drug flunarizine is a good illustration of this overlapping pharmacology, as it can indeed be indicated as vasodilator or anticonvulsant (`https://www.ebi.ac.uk/chembl/ftc/agent/DB04841`), because it interacts non-specifically with either of these channels and can trigger the two response types (Van Nueten et al., 1978).

The corresponding ATC category is C08, the *Calcium channel blockers*. Some repositioning hypotheses have been extracted for this category and are displayed in Table 4.5.

Identifier reference drug	**ATC code reference drug**	**Identifier hypothesis drug**	**ATC code hypothesis drug**	**Hypothesis interpretation**
DB00393 (Nimodipine)	C08 (Calcium channel blockers)	DB00379 (Mexiletine)	C01 (Cardiac therapy)	Similar therapeutic categories (potential effect)
DB00661 (Verapamil)	”	DB00308 (Ibutilide)	C01 (Cardiac therapy)	Similar therapeutic categories (potential effect)
DB01115 (Nifedipine)	”	DB01429 (Aprindine)	C01 (Cardiac therapy)	Similar therapeutic categories (potential effect)
”	”	DB00527 (Dibucaine)	S01 (Ophthalmologicals) D04 (Antipruritics) C05 (Vasoprotectives) S02 (Otologicals) N01 (Anesthetics)	Same MoA, different anatomical location

Table 4.5: Drug repositioning hypotheses related calcium channel blockers (ATC code C08). The symbol ” refers to the value of the previous cell.

Mexiletine, ibutilide and aprindine are all used in generic cardiac therapy and arrhythmia; it is therefore not so surprising to find these drugs as hypothesised calcium channel blockers. More interestingly, dibucaine is present in the list. This drug is traditionally used as an anaesthetic and inhibitor of the sodium pump. Because of these closely related mechanisms of action, the compound is also sometimes used as a vasoprotective, to prevent haemorrhoids for instance. This alternative pharmacology provides evidence about the potential role of this drug for hypertension, which can be further investigated.

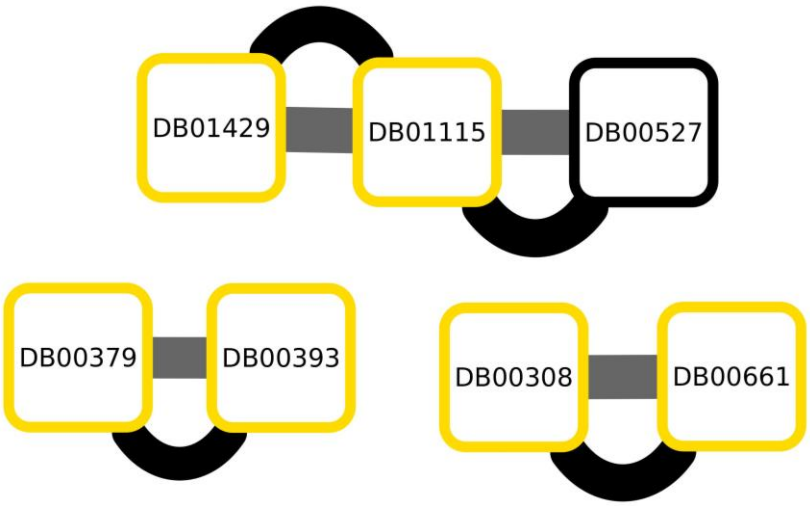


Figure 4.17: Drug repositioning hypotheses related to the ATC category C08 - *Calcium channel blockers*. The width of grey edges is proportional to the functional similarity values between drugs. The black lines represent the structural similarities. Legend for colours and drug name are available online and on the Table 4.5.

4.3.2.3 Adrenergic receptor antagonists

Lowering blood pressure can also be done by preventing the signal transmission mediated by adrenergic receptors, although such agents are principally used to regulate cardiac dysrhythmia. There are different subtypes of adrenergic receptor antagonists, listed as subclasses of the FTC category *Anti-adrenergic receptor activity agent* (FTC_A0004935). The FTC provides a granular representation of this system, first by making the distinction between alpha and beta receptors, and additionally by differentiating the subtype numbers (alpha1 and 2, beta1, 2 and 3).

The corresponding ATC class for this MoA is C07 or *Beta blocking agents* and does not offer such a detailed view on the pharmacological action. Repurposing hypotheses are shown in Table 4.6 and Figure 4.18.

Some of the listed compounds have an already-reported antihypertensive effect, such as amiodarone, dobutamine and arbutamine, and are more generally used in cardiac therapies. Some other molecules are bronchodilators (group B on Table 4.6); they act on the adrenergic receptors expressed in a different location, yet depending on their delivery mechanism they could be potentially active as

Identifier reference drug	**ATC code reference drug**	**Identifier hypothesis drug**	**ATC code hypothesis drug**	**Hypothesis interpretation**
Beta1 adrenergic antagonists (group A)	C07 (Beta blocking agent)	DB01210 (Levobunolol)	S01 (Ophthalmologicals)	Same MoA, different anatomical location
”	”	DB01214 (Metipranolol)	”	”
DB00489 (Sotalol)	”	DB01118 (Amiodarone)	C01 (Cardiac therapy)	Similar therapeutic categories (known effect)
Beta1 adrenergic antagonists (group C)	”	Bronchodilators (group B)	R03 (Drugs for obstructive airway diseases)	Same MoA, different anatomical location
”	”	DB00841 (Dobutamine)	C01 (Cardiac therapy)	Similar therapeutic categories (known effect)
”	”	DB01102 (Arbutamine)	”	Similar therapeutic categories (known effect)
”	”	DB01288 (Fenoterol)	G02 (Other gynaecologicals) R03 (Drugs for obstructive airway diseases)	Same MoA, different anatomical location

Table 4.6: Drug repositioning hypotheses related *Beat-blocking agents* (ATC code C07). The symbol ” refers to the value of the previous cell.

antihypertensive agents. A couple of drugs (metipranolol and levobunolol) are ophthalmological compounds for the treatment of glaucoma (increased pressure in the eye). These compounds are valid hypotheses for the treatment of hypertension. Here again, the common theme behind the repositioning opportunities is the differences in anatomical locations. The same MoA elicited in a different part of the body can totally change the effect of the drug.

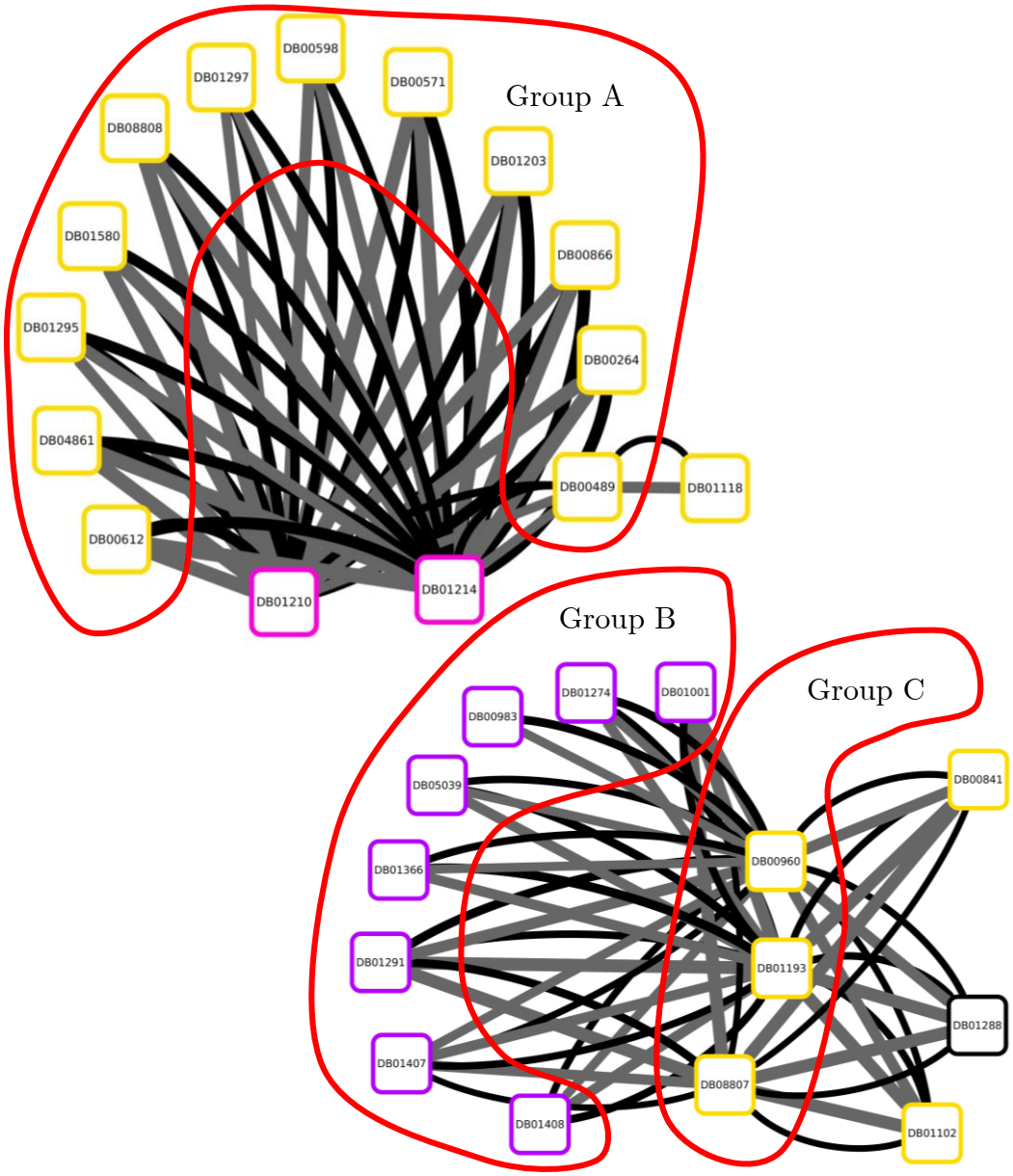


Figure 4.18: Drug repositioning hypotheses related to the ATC category C07 - *Beta blocking agents*. The width of grey edges is proportional to the functional similarity values between drugs. The black lines represent the structural similarities. Legend for colours and drug name are available online and on the Table 4.6.

4.3.2.4 Alpha-2 agonists

Acting on alpha2 receptors can lower blood pressure. A stimulation of these receptors, located in the brain, opens the peripheral arteries, which results in better blood circulation throughout the body (Swedberg et al., 2005). The FTC

category describing this action is *Pro-alpha2-adrenergic receptor activity agent* (FTC_P0004938). The 31 drugs classified as such in the FTC have a variety of clinical applications, from sedatives and analgesics to the expected antihypertensive agents (only 6). No ATC category is dedicated to this mechanism of action, which shows the limit of the gold standard. Even if alpha2 agonists are not the preferred and most common way to reduce hypertension (Nelson, 2010), it is still relevant for drug repositioning purposes to have access to the list of drugs that could act on the receptors, as shown in Table 4.7.

Action	**Number of drugs**
Antihypertensive	6
Dopamine agonist	9
Nasal Decongestants (vasoconstrictor)	6
Ophthalmologicals	2
Analgesics	3
Bronchodilator Agents	2
Other	3

Table 4.7: Actions and number of drugs classified as *Pro-alpha2-adrenergic receptor activity agent* (FTC_P0004938).

Note that the therapeutic groups to be repurposed belong to several categories that have been previously discussed, such as ophthalmological preparations and bronchodilators, but also to new categories such as dopaminergics, which are notably involved in Alzheimer's disease treatment.

4.3.2.5 Diuretics

Diuretic agents promote the production of urine and help the kidneys to eliminate excess salt and water from the blood. This class of compounds also exhibits vasodilating properties; the exact mechanism of the action is still unclear and not directly related to the salt effect.

The FTC describes at least two mechanisms of action responsible for a diuretic effect: *Anti-sodium:potassium:chloride symporter activity agent* (FTC_A0008511 - 9 drugs) and *Anti-sodium:chloride symporter activity agent* (FTC_A0015378 - 7 drugs). All the drugs classified within these categories are all known to be antihypertensive or diuretic agents, without exceptions. The ATC equivalent category is C03 (*diuretics*) which can serve to filter repositioning hypotheses via a

similarity analysis. Table 4.8 and Figure 4.19 provide a summary of repositioning opportunities related to diuretics.

Identifier reference drug	**ATC code reference drug**	**Identifier hypothesis drug**	**ATC code hypothesis drug**	**Hypothesis interpretation**
DB00421 (Spironolactone)	C04 (Diuretics)	DB00665 (Nilutamide)	L02 (Antineoplastics)	Shared off-target
”	”	DB00499 (Flutamide)	L02 (Antineoplastics)	”
”	”	DB01128 (Bicalutamide)	L02 (Antineoplastics)	”
”	”	DB04839 (Cyproterone)	G03 (Sex hormones and modulators of the genital system)	”

Table 4.8: Drug repositioning hypotheses related to *Diuretics* (ATC code C03). The symbol ” refers to the value of the previous cell.

Three drugs (nilutamide, flutamide and bicalutamide) are currently classified as antineoplastic agents and deemed to be also active as antihypertensive via diuretic effect. These predictions are based on the binding of the drugs to the androgen receptor, which has been shown to be tightly linked to the renin-angiotensin system behind the diuretic activity (Ikeda et al., 2009) (Hoshino et al., 2011). Moreover this receptor is an off-target of spironolactone, the known diuretic drug in the pairs (see Table 4.8). The last prediction is cyproterone, which is also an androgen receptor antagonist and is currently used for the treatment of hypersexuality and prostatic carcinoma. Its diuretic activity can be explained in the same manner as the other predictions.

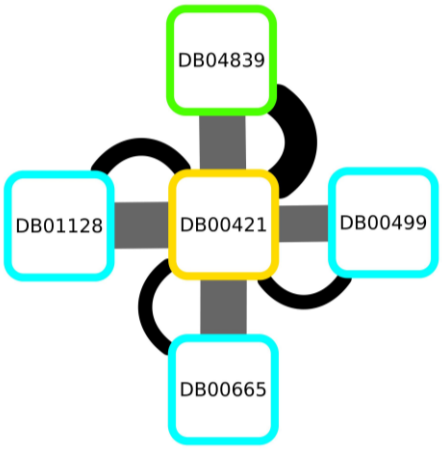


Figure 4.19: Drug repositioning hypotheses related to the ATC category C03 - *Diuretics*. The width of grey edges is proportional to the functional similarity values between drugs. The black lines represent the structural similarities. Legend for colours and drug name are available online and on the Table 4.8.

4.3.2.6 Renin-angiotensin system

The control of salt and water levels in the blood is accomplished by a pathway known as the renin-angiotensin system (see Figure 4.20).

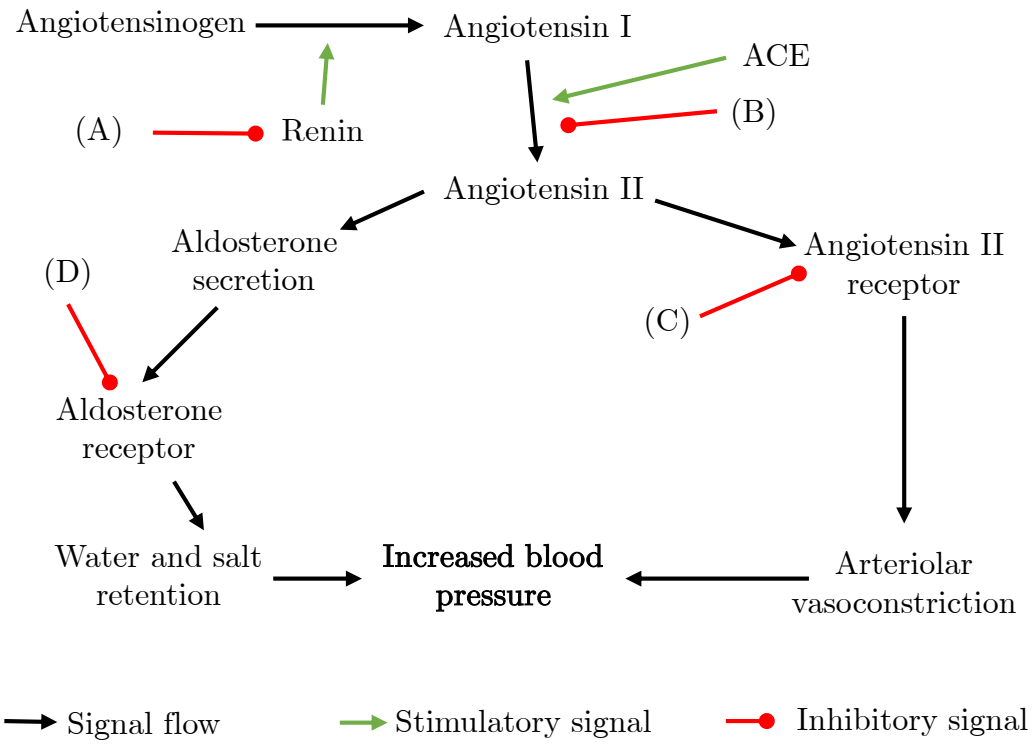


Figure 4.20: Simplified summary of the renin-angiotensin system and its implication in blood pressure regulation. Therapeutic intervention points are indicated with a letter. (A) Renin inhibitors: preventing the synthesis of renin reduces the blood pressure. (B) Angiotensin-converting enzyme (ACE) inhibitors: inhibiting the enzymatic activity prevents the synthesis of angiotensin II and decreases the blood pressure. (C) Angiotensin II receptor agonists: blocks the action of the angiotensin II and reduces blood pressure. (D) Aldosterone receptor antagonists: aldosterone signalling promotes water and salt retention and increases the blood pressure. Preventing its action help to release hypertension.

Angiotensin-converting enzyme (ACE) inhibitors

The angiotensin-converting enzyme (ACE) is the primary target of a series of antihypertensive agents. The enzyme converts angiotensin I into II, a molecule with vasoconstricting properties. Inhibiting the enzymatic activity therefore prevents vasoconstriction and reduces blood pressure (see Figure 4.20-B). The FTC shows limitations in capturing this specific activity. The closest category is *Anti-peptidyl-dipeptidase activity agent* (FTC_A0008241 - 14 drugs), describing the generic chemical catalysis performed by the enzyme. Despite being broadly defined, the 14 drugs present in the class are all antihypertensive agents and manually classified as ACE inhibitors in DrugBank. The ATC better defines the

mechanism of action, as showed by the class C09AA (*ACE inhibitors, plain*). Because of the specificity of the pharmacology, no drug repositioning hypotheses were analysed for this mechanism of action, yet the action is captured in the FTC and usable for further exploration.

Angiotensin II receptor antagonists

The binding of the angiotensin II molecule to receptors triggers arteriolar vasoconstriction. As shown with ACE inhibition in the previous section, preventing angiotensin II synthesis can promote vasodilatation. Another mechanism to lower blood pressure consists of preventing the binding of angiotensin II to its receptor using an antagonist or *blocker* (see Figure 4.20-C).

The FTC contains a category directly capturing this MoA: *Anti-angiotensin type II receptor activity agent* (FTC_A0004945 - 11 drugs). Here again all the drugs are indexed as antihypertensive agents and more precisely as angiotensin II receptor antagonists in DrugBank and in the literature. The category does not contain any unexpected results, therefore no repositioning hypotheses can be advanced using this FTC category alone. The ATC contains a detailed representation of the MoA of interest as well, namely the category C09C or *angiotensin II antagonists, plain*.

Renin inhibitors

Renin is a protein present upstream of the vasoconstriction cascade. In order to be physiologically active and elicit vasoconstriction, renin needs first to be synthesised and produced. Inhibiting its production or modification to its active form therefore prevents vasoconstriction signalling and results in decreased blood pressure (see Figure 4.20-A).

One FTC class only provides information about renin: *Anti-renin secretion into blood stream agent* (FTC_A0002001). This class does not capture the exact MoA by which vasoconstriction is prevented, and no approved drugs have been classified inside this category. This result highlights a limit of the FTC, i.e. more obscure and less common MoAs are not necessarily described in the resource. On the contrary, the ATC features a class for this concept: C09XA, the *renin-inhibitors*.

Aldosterone receptor antagonists

Aldosterone is a hormone secreted by the adrenal gland, located in the cortex. The molecule is involved in signalling which positively regulates blood pressure. Preventing its activity decreases blood pressure; consequently aldosterone receptor antagonists can improve hypertension (see Figure 4.20-D).

This MoA is less traditional than the ones introduced previously in the treatment of hypertension; nonetheless there exist some approved drugs relying on it and described in the ATC: C03DA *Aldosterone antagonists* (4 drugs). Unfortunately such a resolution is not available in the FTC. Two categories merely describe bioprocesses in which aldosterone is involved: *Anti-aldosterone metabolic process agent* (FTC_A0032341) and *Anti-aldosterone secretion agent* (FTC_A0035932), which do not capture the exact MoA nor contain any drugs. This MoA shows again a limitation of the FTC for less frequently used therapeutic categories, as in the previous section.

4.3.2.7 Peripheral vasodilators

Vasodilators are most commonly prescribed for heart problems, but they are also used to treat conditions affecting the blood vessels located in peripheral parts of the body, such as the arms and legs. Raynaud's disease is an example of a condition where peripheral arteries face difficulties to maintain the blood supply, resulting in a numb feeling in the patient when exposed to stress or cold temperatures.

The FTC does not contain a category directly related to this type of action. Indeed, the anatomical location of the pharmacological effect is not covered in the classification; it is therefore impossible to distinguish a peripheral vasodilation from a generic vasodilatation event. On the contrary, the ATC contains a category describing the *peripheral vasodilators* (C04), which can serve to investigate drug repositioning hypotheses (`https://www.ebi.ac.uk/chembl/research/ftc-hypotheses/code/C04`). The generated hypotheses are listed in Table 4.9 and shown in Figure 4.21.

It appears that most of the repositioning hypotheses derived for this group involve molecules used for neurological disorders. First, the series of psychoanaleptics (arousing agent) is already known to produce a range of vascular effects, giving confidence in the validity of the predictions (Khalifa, 2003). A couple

Identifier reference drug	**ATC code reference drug**	**Identifier hypothesis drug**	**ATC code hypothesis drug**	**Hypothesis interpretation**
Drugs used as peripheral vasodilators, group A	C04 (Peripheral vasodilators)	DB06148 (Mianserin)	N06 Psychoanaleptics (Arousing effect)	Known to produce a range of vascular effect according to drugbank
”	”	DB00656 (Trazodone)	”	”
”	”	DB00370 (Mirtazapine)	”	”
”	”	DB01149 (Nefazodone)	”	”
”	”	DB01624 (Zuclopenthixol)	N05 Psycholeptics (Calming effect)	Adrenergic blockade reported from DrugBank, so possible effect on cardiovascular system
”	”	DB01608 (Propericiazine)	”	”
”	”	DB00800 (Fenoldopam)	C01 Cardiac therapy	Same group, known similar action

Table 4.9: Drug repositioning hypotheses related to *peripheral vasodilators* (ATC code C04). The symbol ” refers to the value of the previous cell.

of compounds are psycholeptics (calming agents). The mechanism of action of these drugs relates to adrenergic blockade (Khalifa, 2003); knowing the involvement of the adrenergic receptors in vasoconstriction (cf previous sections), these two psycholeptic agents can plausibly have an effect on peripheral arteries. The last prediction concerns fenoldopam, a drug used in cardiac therapy. The original indication is functionally close to the new predicted one; an activity in the peripheral blood stream is therefore likely to be observed.

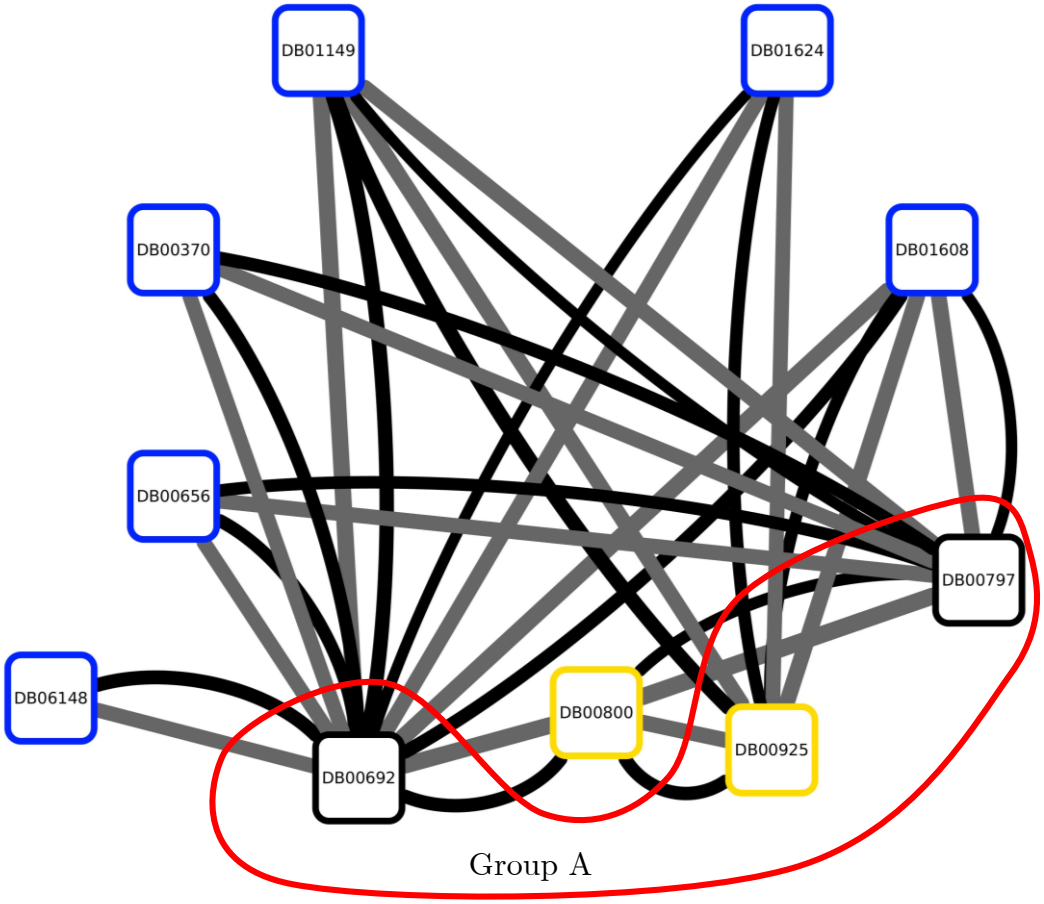


Figure 4.21: Drug repositioning hypotheses related to the ATC category C04 - *Peripheral vasodilators*. The width of grey edges is proportional to the functional similarity values between drugs. The black lines represent the structural similarities. Legend for colours and drug name are available online and on the Table 4.9.

The repositioning hypotheses generated for this group highlight the close relation between neurologicals and vasodilators. Two of the compounds clinically used as peripheral vasodilators (phentolamine and tolazoline) are also already used to treat muscular and joint pains. From this observation, it appears legitimate for other neurological agents to be active in cardiovascular vasodilation.

As a side note for future work, predictions involving neurological drugs are difficult to validate experimentally. The commercial distribution of such molecules is indeed rigidly regulated in most countries, as this class of compounds can also be consumed for illicit recreational use.

The treatment of cardiac hypertension can be addressed from various concep-

tual angles, as presented in this section. For some cases, the FTC describes these abstract MoAs; in such scenarios, the drugs classified in the corresponding FTC category are candidates capable of triggering the given pharmacological effect. This methodology is the toolbox approach, and was not directly used here to derive drug repositioning hypotheses, yet it will be presented in a later section on Alzheimer's disease. On the other hand, some of the conceptually sought MoAs are present in the categories of the gold standard ATC. In such cases repositioning hypotheses can be investigated using the similarity-based approach, as shown in this section using the hypotheses previously generated and available through the web application (see section 4.3).

The drug repositioning hypotheses predicted for cardiovascular hypertension are rooted in similar bioactivity but different anatomical location. For instance, drugs currently indicated for glaucoma, pulmonary problems or against nasal congestion were all listed in the repositioning predictions. These molecules all potentially exhibit vasoconstrictive properties and can further be investigated for the predicted new indication.

4.3.3 Repositioning for Alzheimer's disease

The content of the FTC and the open repositioning hypotheses derived from it can be applied to a wide variety of human diseases. Previous sections exemplified an analysis on vascular hypertension and on the repurposing hypotheses derived using similarity-based methods. In this section, I wanted to show to the reader how the toolbox approach can also provide valuable hypotheses for a dementia currently with no cure: Alzheimer's disease. The technique presented here makes use of the FTC categories as analogues to compartments of a toolbox, helping to find drugs to address the condition, as motivated in Chapter 2.

Alzheimer's disease is a neurodegenerative dementia with a relatively poorly understood biology (McKhann et al., 2011). Phenotypic manifestations include progressive appearance of plaques and tangles in the brain. On the molecular level, several hypotheses are currently under investigation (McKhann et al., 2011). Briefly, two macromolecules appear to be strongly involved in the disease: beta-amyloid peptides, the main component of some of the deposits found in patients' brains, and the Tau protein, responsible for stabilising microtubules (Scheuner et al., 1996) (Delacourte and Defossez, 1986). Given this summary knowledge on

the topic, the goal of the following analysis is to extract from the FTC the MoAs directly related to these two proteins, in order to subsequently identify approved drugs capable of perturbing the dementia.

Five FTC categories containing drugs are directly related to the biological processes of the neurodegenerative condition: *Anti-amyloid precursor protein-biosynthetic-process agent* (FTC_A0042983), *Anti-Tau protein kinase-activity agent* (FTC_A0050321), *Anti-Tau protein binding agent* (FTC_A0048156), *Anti-beta amyloid binding agent* (FTC_A0001540) and finally *Pro-beta amyloid binding agent* (FTC_P0001540). I considered the drugs present inside each of these classes as potential candidates to be repurposed. Table 4.10 shows these drugs, which have been further manually grouped based on the overall similarity of their actions (numbers in Table 4.10).

Within group A, subgroups 1, 2 and 3 are inhibitors of the cholinergic system and some of them, such as galantamine (DB00674), have already been investigated to treat Alzheimer's disease and other related dementias. This class of agent is in line with the cholinergic hypothesis (Babic, 1999), stating that Alzheimer's disease could be caused by dysfunctions in the processing of the acetylcholine.

Subgroup 4 is exclusively composed of barbiturates (central nervous system depressants). The presence of this pharmacological class of compounds as an Alzheimer's disease treatment is more surprising, as very little literature reports on it. Further investigations reveal that the neuronal acetylcholine receptor subunit alpha-7, a common off-target of barbiturates, binds beta-amyloids with high affinity (Wang et al., 2000). As beta-amyloids are themselves strongly involved in the pathology, barbiturates could affect the state of the condition, similarly to cholinergic inhibitors (Wang et al., 2000).

Group B contains four compounds; nicotine and varenicline have been further grouped together because of similar pharmacology (subgroup 5). Nicotine has been shown to improve some of the symptoms of Alzheimer's disease (Jones et al., 1992); it is therefore expected to find this molecule in the predictions. Varenicline possesses a pharmacology related to that of nicotine, which would explain the presence of the drug in this category.

The two remaining drugs of group B are pralidoxime and dipivefrin. Little information is available regarding their potential action against the condition, yet these compounds seem linked to the cholinergic hypothesis and could be considered for experimental testing.

(A) Anti-beta-amyloid binding agent	(B) Pro-beta-amyloid binding agent	(C) Anti-tau protein binding agent	(D) Anti-tau protein kinase activity agent	(E) Anti-amyloid precursor protein biosynthetic process agent
Subgroup 1	Pralidoxime	Vorinostat	Lithium	Hesperetin
Malathion	Dipivefrin			Ezetimibe
Echothiopathe	**Subgroup 5**			
Hexafluronium	Varenicline			
Subgroup 2	Nicotine			
Gallamine				
Triethiodide				
Tubocurarine				
Subgroup 3				
Isoflurophate				
Tacrine				
Edrophonium				
Galantamine				
Ambenonium				
Demecarium				
Pyridostigmine				
Rivastigmine				
Donepezil				
Physostigmine				
Neostigmine				
Subgroup 4				
Amobarbital				
Pentobarbital				
Phenobarbital				
Thiopental				
Talbutal				
Butabarbital				
Primidone				
Metharbital				
Butethal				
Heptabarbital				
Aprobarbital				
Hexobarbital				
Secobarbital				
Butalbital				
Metharbital				
Mehylphenobarbital				

Table 4.10: FTC categories describing some of the modes of action that could impact Alzheimer's disease (letters on figure). The categories have been manually picked on the basis that they could directly affect the dementia. Drugs classified in these FTC categories further manually grouped based on their MoAs similarities (numbers on Figure).

Groups C and D contain one molecule each. These compounds have been classified as agents perturbing some of the physiological function of the Tau protein, a key actor in Alzheimer's disease (Grundke-Iqbal et al., 1986). Vorinostat (group C) is currently indicated for the treatment of cutaneous manifestations in patients with T-cell lymphoma, yet a study has shown *in vivo* (mouse model) the potential of the drug and other histone deacetylase inhibitors in regards to memory deficit (Kilgore et al., 2010). The presence of lithium (group D) was confirmed by a recent study demonstrating a long-term protective effect for the ion in regards to Alzheimer's disease (Young, 2011).

The last group, group E, contains ezetimibe and hesperetin. These two compounds are primarily used as cholesterol lowering agents (statins). As cholesterol metabolism in the brain appears to be related to dementia, statins are believed to prevent or improve the symptoms of the patients. Although early studies (Wolozin, 2004) have failed to clearly show a beneficial effect, a recent clinical trial of over 58,000 patients revealed the positive effect of high potency statins to prevent dementia (ESC13, 2014). Interestingly, this prediction was featured in the FTC before the announcement of the results of the clinical trial.

From the examples briefly presented above, reported and confirmed by the literature, the FTC appears to be suitable to identify real repurposing hypotheses tailored to a disease. Correctly identifying MoAs of interest helps to retrieve the compounds which might impact the treatment of a condition.

4.4 Summary

The FTC was developed to address drug repositioning or more generally indication discovery for active molecules. In this chapter, I illustrated the capability of the resource to fulfil its expectations. First, I discussed the relationship between the similarity of the molecular structure, indication and MoA of approved drugs. Within the dataset considered, the functional and structural descriptors chosen can serve as proxies for the similar property principle: on average, drugs prescribed for increasingly specific indications have increasingly similar molecular structures and functions. More importantly, these descriptors can also be used to identify outliers to the rule, in other words, drugs that are functionally similar yet clinically used for different indications. This set of drug pairs was defined as

the drug repositioning opportunities. The list is openly available online via a web application and serves to investigate the relationship between therapeutic areas (cf section 4.2.4).

The hypotheses cover a wide spectrum of clinical areas, enabling systematic analyses to be performed. However, it is clear that some biological interpretation is still needed on the top of the predictions, in order to understand the molecular reasons behind the MoAs and to move towards laboratory experiments. I therefore decided to focus my attention on two disparate biological dysfunctions: hypertension and Alzheimer's dementia.

From the state-of-the-art presented in Chapter 1, I believe that functional genomics and gene expression experiments are the computational methods that currently produce the most promising results for drug repositioning. I envisioned that this success came from the capacity of these techniques to systematically handle the concept of *function*. In this regard, I specified in Chapter 2 one theoretical approach to studying and handling biomedical knowledge for drug discovery, where organisms are compared to a black box machine to be fixed. I suggested description logics (DLs) as a helpful mathematical framework to address such a task, and in particular the scalable $\mathcal{EL}^{++}$ family. The theory introduced in Chapter 2 found an implementation in chapter 3 with the FTC. This classification defines the mode and mechanism of actions of a panel of approved drugs and is evaluated against the ATC, the traditionally used standard. Finally, an analysis was performed over the data, leading to the large-scale generation of drug repositioning hypotheses presented in this chapter. The work done during my thesis covers only the very beginning of the long hurdled road towards innovative medicine. Knowing that a new drug takes around 12 years and a billion dollars in investment to reach the market (section 1.1), it would have been unrealistic to think that such a goal would have been reached with the time and money allocated for my work. Nonetheless it is possible and important to set the outcomes of the work in the overall drug discovery process, and to discuss the logical next steps which can be undertaken from the results obtained. This will be covered in the final chapter.

4.5 Methods

4.5.1 Structural fingerprints calculation

The fingerprint of a molecule is a simplified - or hashed - representation of its structure, namely a one-dimensional array of 0's and 1's (bits). The presence of a 1 in the fingerprint sequence in a particular position means that a chemical group of interest is present in the molecule, while a 0 means that the group is absent. Four different implementations of fingerprinting were tested. All of them are featured in the Chemistry Development Kit (CDK). I used the version 1.4.19, downloaded from `http://sourceforge.net/projects/cdk/`. The various methods have different rationales, as documented online (`http://cdk.github.io/cdk/1.4/docs/api/`).

- Hybridization fingerprint: default fingerprinting methodology. Does not take into account aromaticity, instead focuses on SP2 molecular hybridisation.

- Extended fingerprint: same as previous, provides support for rings features.

- MACCS fingerprint: generates 166 bit MACCS keys (Dalke-Scientific, 2014). From a series of SMART patterns, the functions output the presence or absence of particular chemical features.

- PubChem fingerprint: generates a large fingerprint (881 bits) from a list of curated rules available online (PubChem, 2014).

4.5.2 Structural similarity between drugs

The relative structural similarity between a pair of drugs was calculated as the Tanimoto similarity value (Tanimoto, 1958) of their fingerprints. Note that Tanimoto similarity is strongly related to the Jaccard index used to calculate semantic similarity. It is defined in section 2.4.3. The similarity value ranges between 0 (totally dissimilar) and 1 (identical) and was calculated using the CDK.

4.5.3 Agreement between fingerprinting methodologies

The respective agreement between the various fingerprinting methodologies listed in section 4.2.1 was defined as the value of the Pearson product-moment correlation coefficient. This coefficient *is a measure of the linear correlation (dependence) between two variables X and Y, giving a value between +1 and 1 inclusive, where 1 is total positive correlation, 0 is no correlation, and 1 is negative correlation* (Häne et al., 1993). In this case, variable X is the structural similarity values between the fingerprints generated with the same method for all drugs (e.g. all pairwise comparisons of fingerprints generated using the MACCS method). The variable Y has the same content, but the values are calculated using a different method. The value of the coefficient is interpreted as: 0 means total disagreement and 1 total agreement between the two methodologies tested.

As four different methods were tested against all the others, each one of them has three coefficient values, representing the different agreement scores (content of a row in Table 4.1). The average of these three values was defined as the overall agreement of one method against other methodologies. The higher the average coefficient, the more one method agrees with the others overall.

4.5.4 Kernel density plots

The kernel density estimates were created using the *stats* package in R. From the documentation of the package (`http://stat.ethz.ch/R-manual/R-patched/library/stats/html/density.html`) the density method *disperses the mass of the empirical distribution function over a regular grid of at least 512 points and then uses the fast Fourier transform to convolve this approximation with a discretized version of the kernel and then uses linear approximation to evaluate the density at the specified points.* The resulting density of similarity can be seen as a value proportional to the chance of drawing from the population a number that is lying in the close proximity of the similarity value. Gaussian smoothing kernel was always used; other parameters were left to default.

4.5.5 Web application related to drug repositioning hypotheses

The source code of the web application is fully open and available at `https://github.com/loopasam/repurposing-analyses`. The application is deployed under `https://www.ebi.ac.uk/chembl/research/ftc-hypotheses`. The project was built using the Play! framework and the Brain library (Croset et al., 2013a), mentioned before in this document. Users can browse the drug repositioning hypotheses in the application, as illustrated in the screen shot Figure 4.22.

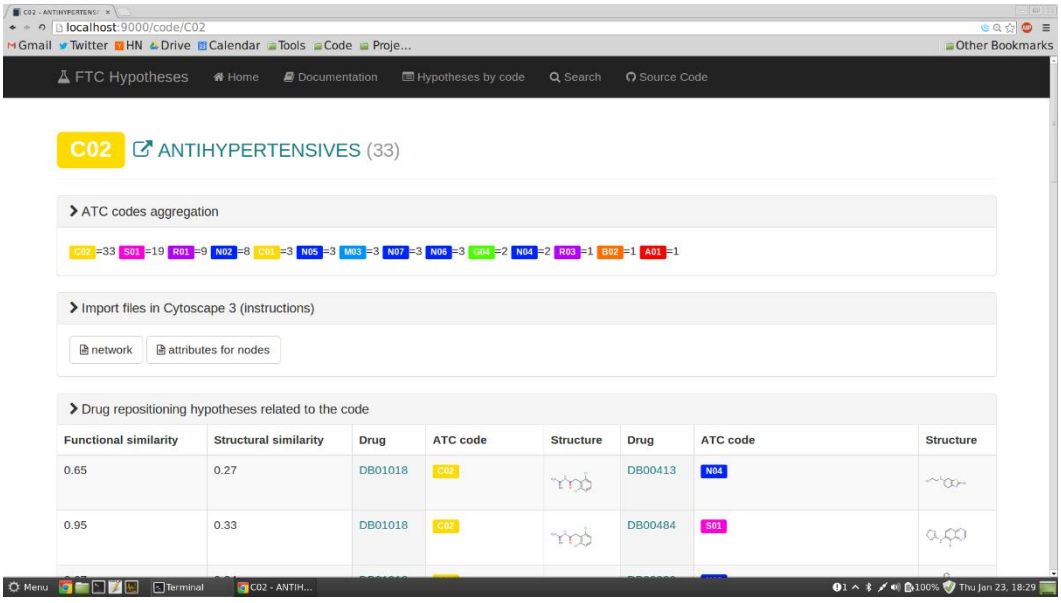


Figure 4.22: Screen shot of the web application to explore drug repositioning hypotheses. The hypotheses are accessible online at `https://www.ebi.ac.uk/chembl/research/ftc-hypotheses`

The hypotheses are accessible from either the ATC code they relate to or by the DrugBank code of the drugs involved in the prediction. The user has the possibility to download the network of predictions as a flat file, for further processing by a tool such as Cytoscape for instance.

4.5.6 Filtering of drug repositioning hypotheses

The list of drug pairs associated in different ATC groups with a resolution of two levels is the starting set used to generate the hypotheses (corresponding to the pairs plotted on Figure 4.5 panel B). I considered the ATC level 2 as a good characterisation of a drug's indication. From this dataset, only the values with a functional similarity superior to 0.6 were conserved. All the other data points were discarded. In this filtered subset, some of the drugs have very few functional annotations; in such cases it is often not clear whether the high functional similarity is biologically meaningful or artefactual and coming from the semantic similarity methodology. In order to filter these ambiguous cases out, I removed all drugs present directly or indirectly in less than 13 FTC classes. The decision is arbitrary and derived from observations and interpretations made over the data. This heuristic aims at first removing the noise, and second discarding drugs about which little is known from a data-centric perspective.

4.5.7 Filtering of relationships between therapeutic areas

The list of drug repositioning hypotheses was used to analyse the links between therapeutic domains. On Figure 4.11 (abstracted - one level), the therapeutic areas have been grouped or abstracted according to their direct ATC super classes. For example, a hypothesis containing drugs indicated for the ATC code A01 and M01 was simplified as a link between the category A and M. The more times a link is present between two ATC groups (first level), the more related these two groups are. When this strength of association was below 25, the association was discarded. This choice is arbitrary and was made in order to remove the noise and extract the most salient features of the dataset.

Figure 4.12 (two levels trimmed) is built from the same data as Figure 4.10, but by removing all the links with a strength inferior to 15 associations. This threshold is also arbitrary and was applied in order to clarify the overall connectivity pattern and look at the most prominent associations.

chapter 5

Outlook and future work

The work presented in this document covered various thematics, all involved in the process of automatically finding new indications for approved drugs. Chapter 1 reviewed the problematic, rationale and drug repositioning efforts that have been done. Chapter 2 argued for a mathematical framework capable of capturing biomedical knowledge: I presented how description logics (DLs) can be used to achieve this task. The formalisation of the molecular machinery was implemented in Chapter 3 via the Functional Therapeutic Chemical Classification System (FTC). The resource is built by integrating information from various repositories using DLs and reasoning services. As a special care was put on scalability, the resource was successfully implemented in practice. Its content and relevance were evaluated against an existing resource, the Anatomical Therapeutic Chemical Classification System (ATC). Chapter 4 describes how the FTC can be used to perform multiple tasks, such as systematically comparing approved drugs based on their function (similarity-based approach) or identifying potential drugs from a known mode of action (MoA, via the toolbox approach). The drug repositioning hypotheses generated cover a wide range of therapeutic areas and are open for the community to investigate via a web application. With this last chapter, I wanted to communicate ideas and future projects that can be performed out of the work accomplished during my thesis and presented in this document.

Firstly, the thesis work covers only a section of the overall drug discovery process. I will present in this chapter the next logical steps and the consider-

ations to account for while moving towards laboratory experiments. Secondly and not directly related to drug repositioning, I will introduce a follow-up analysis comparing the definition of *function* from the FTC perspective against gene expression data. Thirdly and derived from the my personal experience and the theoretical work executed in Chapter 2, I will argue in favour of a simplification of biomedical knowledge representation, based on a straightforward graph structure. Finally, I will briefly describe the clinical relation between off-label use and drug repositioning and how the identification of new indications for a drug can be performed directly from hospital data.

5.1 Work performed in the context of the drug discovery process

Discovering a new active molecule includes an accurate identification of the biological mechanisms behind a disease, followed by the development of chemical compounds screened in assays and further optimised for safety and efficacy. Molecules showing some biological activity in the animal model are then finally tested in clinical trials and approved (Fishman and Porter, 2005) (Cooper, 2002). The approach I presented lies in the beginning of this workflow, where the biology and chemistry are characterised, which usually takes 6 years (see Figure 5.1) and involves numerous experiments. Assuming I worked a maximum of 3 years on the project, I expect at least 3 more years of experimental work in order to gather enough biological evidence and validate some predictions. The fact that the drugs investigated are all already approved should in theory facilitate further development. In practice, drug repositioning can face legal challenges (Chapter 1 section 1.1.2) regarding intellectual protection and safety in particular; therefore it should be assumed that the average time line (see Figure 5.1) will be respected.

I am currently looking to set collaborations with research groups performing experiments and bioactivity assays in order to further explore the hypotheses. As many therapeutic areas are covered, priority to one disease type must be defined first. Hereof, anti-cancer agents are a good starting point. There are 84 hypotheses concerning antineoplastic compounds (`https://www.ebi.ac.uk/chembl/research/ftc-hypotheses/code/L01`), which gives a large enough test case. Phenotypic assays for this class of molecules are well developed and

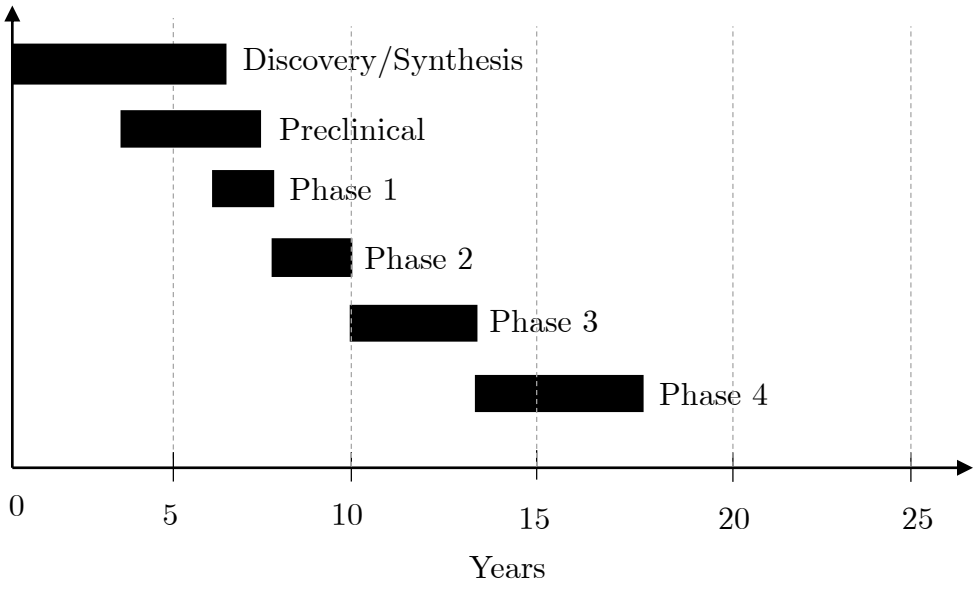


Figure 5.1: Time line of the drug discovery process. My thesis work focused on the Discovery/Synthesis section of the graph.

straightforward: given a line of cancer cells exposed to a chemical, they consist of measuring the ratio of cell death compared to a control (Garnett et al., 2012). The biological context is also better preserved than for target-based assays (Chapter 1 section 1.1.1); this format is therefore more suitable to test the hypotheses, as they have been generated while considering their physiological activities in the human body.

In summary, the work done compares and covers the initial exploration of the path towards a new therapy, which continues at this stage by performing the assays to characterise more accurately the predicted pharmacology.

5.2 Comparison against functional genomics

The FTC introduces one possible representation of the *function* of approved molecules. This descriptor can be used to perform computation, as illustrated throughout this document. The Connectivity Map (Chapter 1 section 1.3.2) defines in another way the biological function of drugs, derived from gene expression experiments. As both methodologies capture the concept of *function* and rely on

similarity-based metrics, it would be interesting to compare the two: given a set of identical drugs, how do the similarity measures behave? Are the data in line with each other? What differences are observed? For which set of compounds? How does the definition of *function* differ from a theoretical perspective (thesis) against the one derived from experimental values (CMap)? As the FTC also provides discrete categories for MoAs (toolbox approach), it could be interesting to investigate what advantage it brings when compared to gene expression data.

In this regard, I have already started to gather the data used in the work done by Iorio et al. (2010). Once available, I can start the investigation and analysis of the definition of approved drugs functions.

5.3 Biomedical knowledge: towards a simpler representation

From a computer science perspective, ontologies are computational artefacts used to represent the knowledge of a domain of interest. The motivation behind the approach is to clarify the semantics of a vocabulary, in order to validate the consistency of a representation and derive implicit information (Chapter 2). The axioms help a computer to unambiguously interpret some knowledge. However, in practice, in the life science domain, human interpretation is always required. For instance, the content of the Gene Ontology is mostly used to perform term enrichment analysis (personal experience) or similarity based computations (Pesquita et al., 2009), in order to guide the interpretation of some generated data. The DLs based representation of the biomedical knowledge presented in Chapter 2 can be arguably seen as heavy and too complicated for the results sought. The axioms present in the FTC are based on existential restriction patterns in order to logically link concepts via an object property. This representation could be simplified by considering all concepts as instances directly linked by a property (section 2.3.3.2) and resulting in a graph. This model is close to the one proposed by the Resource Description Framework (RDF), part of the semantic webs specifications (Klyne and Carroll, 2006). The ambiguity between instances and concepts is known as punning and is allowed by the Web Ontology Language; it is widely used in the RDF representation of biomedical databases such as ChEMBL or Uniprot (Jupp et al., 2014), as it allows for the straightforward representation

of the required information.

Considering this, the future work I will do on biomedical knowledge representation will go in this direction: representing life science knowledge as a simple graph while keeping the expressivity of roles. I expressed most of the FTC semantics using roles, which are highly interoperable one with another (cf section 3.2.9.2). As far as I know, there exists no framework within the semantic web specification stack (RDF, RDFS, OWL and related profiles) allowing the extended use of object properties or role axioms (transitivity, chained property) in combination with a simple graph representation like the one featured by RDF. With such a language, it would be possible to keep the advantages of scalable reasoning (subset of $\mathcal{EL}^{++}$) while considerably simplifying the initial representation and discarding cumbersome axioms not so helpful for biological interpretation later on, for instance existential restrictions.

This language will lie between network biology (Ravasz et al., 2002), semantic nets (Schubert, 1978), Petri Nets (Murata, 1989), pathways databases (Schaefer et al., 2009) and RDF while keeping the important reasoning services. The added value on the top of these previous approaches will reside in the importance given to the edges between nodes, in particular their biomedical semantics and what can be derived and expanded from them. The black box machine analogy is still valid for this representation, and the implementation can be done in many other ways than the one presented with the FTC. The focus of the language would be the deductions that can be derived from it rather than maximizing interoperability between resources. A graph representation also allows full control of data querying and analysis; it would also be easier to derive metrics such as similarity values, commonly used in biology (Stevens et al., 2007). Moreover, some robust solutions are available at the time of writing (Have and Jensen, 2013).

More work has to be done in order to fully express the idea, but it can be summarised as a simple and customisable graph-based framework to locally integrate biomedical information, with a strong focus on edges semantics interoperability (*part-of*, *regulates*, *involved-in* combined with rules).

5.4 Off-label uses and clinical drug repositioning

In the context of my work, I came to discuss drug repositioning with some physicians practising in a hospital (Addenbrooke's - Cambridge UK). After the meetings and analyses presented in Chapter 4 (section 4.3.1), it clearly appeared that drug repositioning relates closely to off-label prescriptions. In practice, doctors indicate a treatment for a different usage than what it is regulated for 20 percent of the time. (Cras et al., 2007). Sometimes a new pharmacological effect is identified (Stephens and Brynner, 2009), and the indication line of the drug can be expanded accordingly. This type of evidence is very valuable, as it reflects the effect of the drug directly in human population. A meticulous recording and automated analysis can help to identify such cases rapidly. Off-label prescription data is currently not recorded in all hospitals (personal discussion with physicians), and a computer system handling it might be a smart way to directly capture relevant hypotheses. It could be developed on the top of an electronic medical records system (Cras et al., 2007) or as a stand-alone application. I am currently investigating what would be the requirements of such a workflow, in collaboration with physicians and biomedical researchers.

This chapter concludes the thesis work I decided to present in more details. The reader can refer to the Appendices in order to find the full list of the peer-reviewed articles I contributed to and published in the course of the doctorate, as a complementary source of information. I hope to have presented clearly and concisely the message I wanted to communicate: living organisms, just as machines, are subject to dysfunction, which can be investigated for repair using description logics. The source of this document is free, open and available online at `https://github.com/loopasam/thesis`.

Bibliography

Allen, J. F. and Frisch, A. M. (1982). What's in a semantic network? In *Proceedings of the 20th annual meeting on Association for Computational Linguistics*, pages 19–27. Association for Computational Linguistics.

An, Y. Y. L. J. and Jones, S. J. (2006). A large-scale computational approach to drug repositioning. *Genome Informatics*, 17(2):239–247.

Antezana, E., Kuiper, M., and Mironov, V. (2009). Biological knowledge management: the emerging role of the semantic web technologies. *Briefings in bioinformatics*, 10(4):392–407.

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

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29.

ATC (January 2014a). A03. `http://www.whocc.no/atc_ddd_index/?code=A03&showdescription=yes`.

ATC (January 2014b). R05. `http://www.whocc.no/atc_ddd_index/?code=R05&showdescription=yes`.

Baader, F., Brandt, S., and Lutz, C. (2005). Pushing the el envelope. In *IJCAI*, volume 5, pages 364–369.

Baader, F., Brandt, S., and Lutz, C. (2008). Pushing the el envelope further.

Babic, T. (1999). The cholinergic hypothesis of alzheimers disease: a review of progress. *Journal of Neurology, Neurosurgery & Psychiatry*, 67(4):558–558.

Barabasi, A.-L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113.

Barratt, M. J. and Frail, D. E. (2012). *Drug repositioning: Bringing new life to shelved assets and existing drugs*. Wiley. com.

Bende, M., Fukami, M., Arfors, K., Mark, J., Stierna, P., and Intaglietta, M. (1996). Effect of oxymetazoline nose drops on acute sinusitis in the rabbit. *The Annals of otology, rhinology, and laryngology*, 105(3):222–225.

Berners-Lee, T., Hendler, J., Lassila, O., et al. (2001). The semantic web. *Scientific american*, 284(5):28–37.

Black, L. J., Sato, M., Rowley, E., Magee, D., Bekele, A., Williams, D., Cullinan, G., Bendele, R., Kauffman, R., Bensch, W., et al. (1994). Raloxifene (ly139481 hci) prevents bone loss and reduces serum cholesterol without causing uterine hypertrophy in ovariectomized rats. *Journal of Clinical Investigation*, 93(1):63.

Bodenreider, O. (2004). The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270.

Bonica, J. et al. (1979). The need of a taxonomy. *Pain*, 6(3):247–248.

Brooksbank, C., Bergman, M. T., Apweiler, R., Birney, E., and Thornton, J. (2014). The european bioinformatics institutes data resources 2014. *Nucleic acids research*, 42(D1):D18–D25.

Campillos, M., Kuhn, M., Gavin, A.-C., Jensen, L. J., and Bork, P. (2008). Drug target identification using side-effect similarity. *Science*, 321(5886):263–266.

Chiang, A. P. and Butte, A. J. (2009). Systematic evaluation of drug–disease relationships to identify leads for novel drug uses. *Clinical Pharmacology & Therapeutics*, 86(5):507–510.

Cobham, A. (1965). The intrinsic computational difficulty of functions. In *Proceedings of the 1964 Congress for Logic, Methodology, and the Philosophy of Science*, pages 24–30.

Consortium, U. et al. (2013). Update on activities at the universal protein resource (uniprot) in 2013. *Nucleic acids research*, 41(D1):D43–D47.

Cooper, M. A. (2002). Optical biosensors in drug discovery. *Nature Reviews Drug Discovery*, 1(7):515–528.

Cras, A., Conscience, M.-A., Rajzbaum, G., Lillo-Le Louët, A., Lopez, N., Tersen, I., and Bezie, Y. (2007). Off-label prescribing in a french hospital. *Pharmacy world & science*, 29(2):97–100.

Croset, S., Overington, J. P., and Rebholz-Schuhmann, D. (2013a). Brain: biomedical knowledge manipulation. *Bioinformatics*, 29(9):1238–1239.

Croset, S., Overington, J. P., and Rebholz-Schuhmann, D. (2013b). The functional therapeutic chemical classification system. *Bioinformatics*, page btt628.

Dalke-Scientific (January 2014). Fingerprints. `http://www.dalkescientific.com/writings/NBN/fingerprints.html`.

Darwin, C. and Bynum, W. F. (2009). *The origin of species by means of natural selection: or, the preservation of favored races in the struggle for life*. AL Burt.

De Franchi, E., Schalon, C., Messa, M., Onofri, F., Benfenati, F., and Rognan, D. (2010). Binding of protein kinase inhibitors to synapsin i inferred from pair-wise binding site similarity measurements. *PLoS One*, 5(8):e12214.

Delacourte, A. and Defossez, A. (1986). Alzheimer's disease: tau proteins, the promoting factors of microtubule assembly, are major components of paired helical filaments. *Journal of the neurological sciences*, 76(2):173–186.

DiMasi, J. A. (2001). New drug development in the united states from 1963 to 1999. *Clinical pharmacology and therapeutics*, 69(5):286.

Dimmer, E. C., Huntley, R. P., Alam-Faruque, Y., Sawford, T., O'Donovan, C., Martin, M. J., Bely, B., Browne, P., Chan, W. M., Eberhardt, R., et al. (2012). The uniprot-go annotation database in 2011. *Nucleic acids research*, 40(D1):D565–D570.

Dreger, A. D. (1998). ambiguous sexor ambivalent medicine?: Ethical issues in the treatment of intersexuality. *Hastings Center Report*, 28(3):24–35.

Dudley, J. T., Deshpande, T., and Butte, A. J. (2011a). Exploiting drug–disease relationships for computational drug repositioning. *Briefings in bioinformatics*, 12(4):303–311.

Dudley, J. T., Sirota, M., Shenoy, M., Pai, R., Roedder, S., Chiang, A. P., Morgan, A. A., Sarwal, M., Pasricha, P. J., and Butte, A. J. (2011b). Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. *Science translational medicine*, 3(96):96ra76.

Duran-Frigola, M., Aloy, P., et al. (2012). Recycling side-effects into clinical markers for drug repositioning. *Genome Med*, 4(3).

Eckert, H. and Bajorath, J. (2007). Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug discovery today*, 12(5):225–233.

Editorial (2013). Form and function. *Nature*, 495(7440):141–142.

ESC13 (January 2014). High dose statins prevent dementia. `http://www.escardio.org/about/press/press-releases/esc13-amsterdam/Pages/statins-dementia-elderly.aspx`.

Evans, J. M., Leonelli, F. M., Ziegler, M. G., McIntosh, C. M., Patwardhan, A. R., Ertl, A. C., Kim, C. S., and Knapp, C. F. (2001). Epinephrine, vasodilation and hemoconcentration in syncopal, healthy men and women. *Autonomic Neuroscience*, 93(1):79–90.

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

Fisher, J. and Henzinger, T. A. (2007). Executable cell biology. *Nature biotechnology*, 25(11):1239–1249.

Fishman, M. C. and Porter, J. A. (2005). Pharmaceuticals: a new grammar for drug discovery. *Nature*, 437(7058):491–493.

Fliri, A. F., Loging, W. T., Thadeio, P. F., and Volkmann, R. A. (2005). Analysis of drug-induced effect patterns to link structure and side effects of medicines. *Nature chemical biology*, 1(7):389–397.

Fliri, A. F., Loging, W. T., and Volkmann, R. A. (2007). Analysis of system structure–function relationships. *ChemMedChem*, 2(12):1774–1782.

Franke, A., McGovern, D. P., Barrett, J. C., Wang, K., Radford-Smith, G. L., Ahmad, T., Lees, C. W., Balschun, T., Lee, J., Roberts, R., et al. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed crohn's disease susceptibility loci. *Nature genetics*, 42(12):1118–1125.

Furberg, H., Kim, Y., Dackor, J., Boerwinkle, E., Franceschini, N., Ardissino, D., Bernardinelli, L., Mannucci, P., Mauri, F., Merlini, P. A., et al. (2010). Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nature genetics*, 42(5):441–447.

Garnett, M. J., Edelman, E. J., Heidorn, S. J., Greenman, C. D., Dastur, A., Lau, K. W., Greninger, P., Thompson, I. R., Luo, X., Soares, J., et al. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 483(7391):570–575.

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., et al. (2012). Chembl: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(D1):D1100–D1107.

Gengo, F. and Gabos, C. (1987). Antihistamines, drowsiness, and psychomotor impairment: central nervous system effect of cetirizine. *Annals of allergy*, 59(6 Pt 2):53–57.

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

Gillespie, D. T. (2007). Stochastic simulation of chemical kinetics. *Annu. Rev. Phys. Chem.*, 58:35–55.

GO (January 2014). Go relations. `http://www.geneontology.org/GO.ontology.relations.shtml`.

Golbreich, C., Zhang, S., and Bodenreider, O. (2006). The foundational model of anatomy in owl: Experience and perspectives. *Web Semantics: Science, Services and Agents on the World Wide Web*, 4(3):181–195.

Gonçalves, R. S., Bail, S., Jimenez-Ruiz, E., Matentzoglu, N., Parsia, B., Glimm, B., and Kazakov, Y. (2013). Owl reasoner evaluation (ore) workshop 2013 results: Short report. In *2nd OWL Reasoner Evaluation Workshop (ORE 2013)*, page 1.

Gottlieb, A., Stein, G. Y., Ruppin, E., and Sharan, R. (2011). Predict: a method for inferring novel drug indications with application to personalized medicine. *Molecular systems biology*, 7(1).

Gruber, T., Ontology, I. L. L., and Özsu, M. T. (2009). Encyclopedia of database systems. *Ontology*.

Grundke-Iqbal, I., Iqbal, K., Tung, Y.-C., Quinlan, M., Wisniewski, H. M., and Binder, L. I. (1986). Abnormal phosphorylation of the microtubule-associated protein tau (tau) in alzheimer cytoskeletal pathology. *Proceedings of the National Academy of Sciences*, 83(13):4913–4917.

Hanage, W. P. (2013). Fuzzy species revisited. *BMC biology*, 11(1):41.

Häne, B. G., Jäger, K., and Drexler, H. G. (1993). The pearson product-moment correlation coefficient is better suited for identification of dna fingerprint profiles than band matching algorithms. *Electrophoresis*, 14(1):967–972.

Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999). From molecular to modular cell biology. *Nature*, 402:C47–C52.

Hastings, J., de Matos, P., Dekker, A., Ennis, M., Harsha, B., Kale, N., Muthukrishnan, V., Owen, G., Turner, S., Williams, M., et al. (2013). The chebi reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic acids research*, 41(D1):D456–D463.

Haupt, V. J. and Schroeder, M. (2011). Old friends in new guise: repositioning of known drugs with structural bioinformatics. *Briefings in bioinformatics*, 12(4):312–326.

Have, C. T. and Jensen, L. J. (2013). Are graph databases ready for bioinformatics? *Bioinformatics*, 29(24):3107.

Hendler, J. (January 2014). A little semantics goes a long way. `http://www.cs.rpi.edu/~hendler/LittleSemanticsWeb.html`.

Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P. F., and Rudolph, S. (2009). Owl 2 web ontology language primer. *W3C recommendation*, 27(1):123.

Hoehndorf, R., Dumontier, M., and Gkoutos, G. V. (2012). Identifying aberrant pathways through integrated analysis of knowledge in pharmacogenomics. *Bioinformatics*, 28(16):2169–2175.

Hoehndorf, R., Dumontier, M., Oellrich, A., Wimalaratne, S., Rebholz-Schuhmann, D., Schofield, P., and Gkoutos, G. V. (2011a). A common layer of interoperability for biomedical ontologies based on owl el. *Bioinformatics*, 27(7):1001–1008.

Hoehndorf, R., Loebe, F., Kelso, J., and Herre, H. (2007). Representing default knowledge in biomedical ontologies: Application to the integration of anatomy and phenotype ontologies. *BMC bioinformatics*, 8(1):377.

Hoehndorf, R., Oellrich, A., Dumontier, M., Kelso, J., Rebholz-Schuhmann, D., and Herre, H. (2010). Relations as patterns: bridging the gap between obo and owl. *BMC bioinformatics*, 11(1):441.

Hoehndorf, R., Schofield, P. N., and Gkoutos, G. V. (2011b). Phenomenet: a whole-phenome approach to disease gene discovery. *Nucleic acids research*, 39(18):e119–e119.

Hopkins, A. L. (2008). Network pharmacology: the next paradigm in drug discovery. *Nature chemical biology*, 4(11):682–690.

Horridge, M. and Bechhofer, S. (2011). The owl api: A java api for owl ontologies. *Semantic Web*, 2(1):11–21.

Horridge, M., Drummond, N., Goodwin, J., Rector, A. L., Stevens, R., and Wang, H. (2006). The manchester owl syntax. In *OWLed*, volume 216.

Hoshino, K., Ishiguro, H., Teranishi, J.-i., Yoshida, S.-i., Umemura, S., Kubota, Y., and Uemura, H. (2011). Regulation of androgen receptor expression through angiotensin ii type 1 receptor in prostate cancer cells. *The Prostate*, 71(9):964–975.

Hunskaar, S. and Hole, K. (1987). The formalin test in mice: dissociation between inflammatory and non-inflammatory pain. *Pain*, 30(1):103–114.

Hurle, M., Yang, L., Xie, Q., Rajpal, D., Sanseau, P., and Agarwal, P. (2013). Computational drug repositioning: From data to therapeutics. *Clinical Pharmacology & Therapeutics*.

Ikeda, Y., Aihara, K.-i., Yoshida, S., Sato, T., Yagi, S., Iwase, T., Sumitomo, Y., Ise, T., Ishikawa, K., Azuma, H., et al. (2009). Androgen-androgen receptor system protects against angiotensin ii-induced vascular remodeling. *Endocrinology*, 150(6):2857–2864.

Iorio, F., Bosotti, R., Scacheri, E., Belcastro, V., Mithbaokar, P., Ferriero, R., Murino, L., Tagliaferri, R., Brunetti-Pierri, N., Isacchi, A., et al. (2010). Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proceedings of the National Academy of Sciences*, 107(33):14621–14626.

Jaccard, P. (1912). The distribution of the flora in the alpine zone. 1. *New phytologist*, 11(2):37–50.

Johnson, M. A. and Maggiora, G. M. (1990). *Concepts and applications of molecular similarity*. Wiley New York.

Jones, G., Sahakian, B., Levy, R., Warburton, D. M., and Gray, J. (1992). Effects of acute subcutaneous nicotine on attention, information processing and short-term memory in alzheimer's disease. *Psychopharmacology*, 108(4):485–494.

Jordan, V. C., Phelps, E., and 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.

Jupp, S., Malone, J., Bolleman, J., Brandizi, M., Davies, M., Garcia, L., Gaulton, A., Gehant, S., Laibe, C., Redaschi, N., et al. (2014). The ebi rdf platform: linked open data for the life sciences. *Bioinformatics*, page btt765.

Jupp, S., Stevens, R., and Hoehndorf, R. (2012). Logical gene ontology annotations (goal): exploring gene ontology annotations with owl. *J Biomed Semantics*, 3.

Kathiresan, S., Melander, O., Guiducci, C., Surti, A., Burtt, N. P., Rieder, M. J., Cooper, G. M., Roos, C., Voight, B. F., Havulinna, A. S., et al. (2008). Six

new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nature genetics*, 40(2):189–197.

Kazakov, Y., Krötzsch, M., and Simančík, F. (2013). The incredible elk. *Journal of Automated Reasoning*, pages 1–61.

Keiser, M. J., Setola, V., Irwin, J. J., Laggner, C., Abbas, A. I., Hufeisen, S. J., Jensen, N. H., Kuijer, M. B., Matos, R. C., Tran, T. B., et al. (2009). Predicting new molecular targets for known drugs. *Nature*, 462(7270):175–181.

Khalifa, A. E. (2003). Zuclopenthixol facilitates memory retrieval in rats: possible involvement of noradrenergic and serotonergic mechanisms. *Pharmacology Biochemistry and Behavior*, 75(4):755–762.

Kilgore, M., Miller, C. A., Fass, D. M., Hennig, K. M., Haggarty, S. J., Sweatt, J. D., and Rumbaugh, G. (2010). Inhibitors of class 1 histone deacetylases reverse contextual memory deficits in a mouse model of alzheimer's disease. *Neuropsychopharmacology*, 35(4):870–880.

Kim, S.-K., Song, M.-Y., Kim, C., Yea, S.-J., Jang, H. C., and Lee, K.-C. (2008). Temporal ontology language for representing and reasoning interval-based temporal knowledge. In *The Semantic Web*, pages 31–45. Springer.

Kinnings, S. L., Liu, N., Buchmeier, N., Tonge, P. J., Xie, L., and Bourne, P. E. (2009). Drug discovery using chemical systems biology: repositioning the safe medicine comtan to treat multi-drug and extensively drug resistant tuberculosis. *PLoS computational biology*, 5(7):e1000423.

Klyne, G. and Carroll, J. J. (2006). Resource description framework (rdf): Concepts and abstract syntax.

Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., et al. (2011). Drugbank 3.0: a comprehensive resource for omics research on drugs. *Nucleic acids research*, 39(suppl 1):D1035–D1041.

Knublauch, H., Horridge, M., Musen, M. A., Rector, A. L., Stevens, R., Drummond, N., Lord, P. W., Noy, N. F., Seidenberg, J., and Wang, H. (2005). The protege owl experience. In *OWLED*.

Krall, J., Fittingoff, M., and Rajfer, J. (1988). Characterization of cyclic nucleotide and inositol 1, 4, 5-trisphosphate-sensitive calcium-exchange activity of smooth muscle cells cultured from the human corpora cavernosa. *Biology of reproduction*, 39(4):913–922.

Krötzsch, M. (2012). Owl 2 profiles: An introduction to lightweight ontology languages. In *Reasoning Web. Semantic Technologies for Advanced Query Answering*, pages 112–183. Springer.

Kruger, F. A., Rostom, R., and Overington, J. P. (2012). Mapping small molecule binding data to structural domains. *BMC bioinformatics*, 13(Suppl 17):S11.

Kubinyi, H. (1998). Similarity and dissimilarity: a medicinal chemist's view. *Perspectives in Drug Discovery and Design*, 9:225–252.

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

Kunkel, S. D., Suneja, M., Ebert, S. M., Bongers, K. S., Fox, D. K., Malmberg, S. E., Alipour, F., Shields, R. K., and Adams, C. M. (2011). mrna expression signatures of human skeletal muscle atrophy identify a natural compound that increases muscle mass. *Cell metabolism*, 13(6):627–638.

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., Lerner, J., Brunet, J.-P., Subramanian, A., Ross, K. N., et al. (2006). The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *science*, 313(5795):1929–1935.

Law, M., Wald, N., Morris, J., et al. (2003). Lowering blood pressure to prevent myocardial infarction and stroke: a new preventive strategy.

Lazebnik, Y. (2002). Can a biologist fix a radio?-or, what i learned while studying apoptosis. *Cancer cell*, 2(3):179–182.

Le Novere, N., Bornstein, B., Broicher, A., Courtot, M., Donizelli, M., Dharuri, H., Li, L., Sauro, H., Schilstra, M., Shapiro, B., et al. (2006). Biomodels database: a free, centralized database of curated, published, quantitative

kinetic models of biochemical and cellular systems. *Nucleic acids research*, 34(suppl 1):D689–D691.

Li, Q., Cheng, T., Wang, Y., and Bryant, S. H. (2010). Pubchem as a public resource for drug discovery. *Drug discovery today*, 15(23):1052–1057.

Li, Y. and Agarwal, P. (2009). A pathway-based view of human diseases and disease relationships. *PloS one*, 4(2):e4346.

LINCS (January 2014). Connectivity map. `http://lincscloud.org/`.

Lipinski, C. A., Lombardo, F., Dominy, B. W., and Feeney, P. J. (1997). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews*, 23(1):3–25.

Lord, P. (2013). The semantic web takes wing: Programming ontologies with tawny-owl. *arXiv preprint arXiv:1303.0213*.

Lounkine, E., Keiser, M. J., Whitebread, S., Mikhailov, D., Hamon, J., Jenkins, J. L., Lavan, P., Weber, E., Doak, A. K., Côté, S., et al. (2012). Large-scale prediction and testing of drug activity on side-effect targets. *Nature*, 486(7403):361–367.

Machado, D., Costa, R. S., Rocha, M., Ferreira, E. C., Tidor, B., and Rocha, I. (2011). Modeling formalisms in systems biology. *AMB Express*, 1(1):1–14.

Martin, Y. C., Kofron, J. L., and Traphagen, L. M. (2002). Do structurally similar molecules have similar biological activity? *Journal of medicinal chemistry*, 45(19):4350–4358.

McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack Jr, C. R., Kawas, C. H., Klunk, W. E., Koroshetz, W. J., Manly, J. J., Mayeux, R., et al. (2011). The diagnosis of dementia due to alzheimers disease: Recommendations from the national institute on aging-alzheimers association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimer's & Dementia*, 7(3):263–269.

Medina-Franco, J. L., Giulianotti, M. A., Welmaker, G. S., and Houghten, R. A. (2013). Shifting from the single to the multitarget paradigm in drug discovery. *Drug discovery today*, 18(9):495–501.

Men, A., Le, C., Charlab, R., Gobburu, J., and Lesko, L. (2010). 1: Sardana d, zhu c, zhang m, gudivada rc, yang l, jegga ag. drug repositioning for orphan diseases. brief bioinform. 2011 jul; 12 (4): 346-56. doi. *Policy*, 8(5):343–50.

Mendel, G. (1866). Versuche über pflanzenhybriden. *Verhandlungen des naturforschenden Vereines in Brunn 4: 3*, 44.

Meng, T. C., Somani, S., and Dhar, P. (2004). Modeling and simulation of biological systems with stochasticity. *In silico biology*, 4(3):293–309.

Motik, B., Grau, B. C., Horrocks, I., Wu, Z., Fokoue, A., and Lutz, C. (2009). Owl 2 web ontology language: Profiles. *W3C recommendation*, 27:61.

Mungall, C. J., Bada, M., Berardini, T. Z., Deegan, J., Ireland, A., Harris, M. A., Hill, D. P., and Lomax, J. (2011). Cross-product extensions of the gene ontology. *Journal of biomedical informatics*, 44(1):80–86.

Mungall, C. J., Gkoutos, G. V., Smith, C. L., Haendel, M. A., Lewis, S. E., and Ashburner, M. (2010). Integrating phenotype ontologies across multiple species. *Genome biology*, 11(1):R2.

Murata, T. (1989). Petri nets: Properties, analysis and applications. *Proceedings of the IEEE*, 77(4):541–580.

Napolitano, F., Zhao, Y., Moreira, V. M., Tagliaferri, R., Kere, J., D'Amato, M., and Greco, D. (2013). Drug repositioning: A machine-learning approach through data integration. *Journal of Cheminformatics*, 5(1):30.

Nelson, M. (2010). Drug treatment of elevated blood pressure. *Australian Prescriber*, 33(4).

Nelson, S. J., Schopen, M., Savage, A. G., Schulman, J.-L., and Arluk, N. (2004). The mesh translation maintenance system: structure, interface design, and implementation. *Medinfo*, 11(Pt 1):67–69.

Nettles, J. H., Jenkins, J. L., Bender, A., Deng, Z., Davies, J. W., and Glick, M. (2006). Bridging chemical and biological space:target fishing using 2d and 3d molecular descriptors. *Journal of medicinal chemistry*, 49(23):6802–6810.

Neumann, J. v. and Burks, A. W. (1966). Theory of self-reproducing automata.

New-York-Times-Archives (March 1998). U.s. approves sale of impotence 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`.

NHS-Choices (January 2014a). Angina. `http://www.nhs.uk/conditions/Angina/Pages/Introduction.aspx`.

NHS-Choices (January 2014b). High blood pressure. `http://www.nhs.uk/conditions/Blood-pressure-(high)/Pages/Introduction.aspx`.

Nielsen, F. H. (1999). Ultratrace minerals. *Modern nutrition in health and disease*, 9:283–303.

Noeske, T., Sasse, B. C., Stark, H., Parsons, C. G., Weil, T., and Schneider, G. (2006). Predicting compound selectivity by self-organizing maps: Cross-activities of metabotropic glutamate receptor antagonists. *ChemMedChem*, 1(10):1066–1068.

OBO (January 2014). Obo foundry. `http://www.obofoundry.org/`.

Ong, K. L., Cheung, B. M., Man, Y. B., Lau, C. P., and Lam, K. S. (2007). Prevalence, awareness, treatment, and control of hypertension among united states adults 1999–2004. *Hypertension*, 49(1):69–75.

Organization, W. H. et al. (2006). The anatomical therapeutic chemical classification system with defined daily doses (atc/ddd). *Norway: WHO*.

Paolini, G. V., Shapland, R. H., van Hoorn, W. P., Mason, J. S., and Hopkins, A. L. (2006). Global mapping of pharmacological space. *Nature biotechnology*, 24(7):805–815.

Pesquita, C., Faria, D., Falcao, A. O., Lord, P., and Couto, F. M. (2009). Semantic similarity in biomedical ontologies. *PLoS computational biology*, 5(7):e1000443.

PubChem (January 2014). Pubchem substructure fingerprint. ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.txt.

Quigley, H. A. (1996). Number of people with glaucoma worldwide. *British Journal of Ophthalmology*, 80(5):389–393.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabási, A.-L. (2002). Hierarchical organization of modularity in metabolic networks. *science*, 297(5586):1551–1555.

Renaud, R. C. and Xuereb, H. (2002). Erectile-dysfunction therapies. *Nature Reviews Drug Discovery*, 1(9):663–664.

Rogers, D. J. and Tanimoto, T. T. (1960). A computer program for classifying plants. *Science*, 132(3434):1115–1118.

Sampaio, E. P., Sarno, E. N., Galilly, R., Cohn, Z. A., and Kaplan, G. (1991). Thalidomide selectively inhibits tumor necrosis factor alpha production by stimulated human monocytes. *The Journal of experimental medicine*, 173(3):699–703.

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

Sanseau, P. and Koehler, J. (2011). Editorial: Computational methods for drug repurposing. *Briefings in bioinformatics*, 12(4):301–302.

Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., and Buetow, K. H. (2009). Pid: the pathway interaction database. *Nucleic acids research*, 37(suppl 1):D674–D679.

Scheuner, D., Eckman, C., Jensen, M., Song, X., Citron, M., Suzuki, N., Bird, T., Hardy, J., Hutton, M., Kukull, W., et al. (1996). Secreted amyloid β–protein similar to that in the senile plaques of alzheimer's disease is increased in vivo by the presenilin 1 and 2 and app mutations linked to familial alzheimer's disease. *Nature medicine*, 2(8):864–870.

Schubert, L. K. (1978). *The structure and organization of a semantic net for comprehension and inference*. Department of Computing Science, University of Alberta.

Schulz, S. and Jansen, L. (2013). Formal ontologies in biomedical knowledge representation. *Yearbook of medical informatics*, 8(1):132.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504.

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

Stafford, R. S. (2008). Regulating off-label drug userethinking the role of the fda. *New England Journal of Medicine*, 358(14):1427–1429.

Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E., and Willighagen, E. (2003). The chemistry development kit (cdk): An open-source java library for chemo-and bioinformatics. *Journal of chemical information and computer sciences*, 43(2):493–500.

Stephens, T. D. and Brynner, R. (2009). *Dark remedy: The impact of thalidomide and its revival as a vital medicine*. Basic Books.

Stevens, R., Egana Aranguren, M., Wolstencroft, K., Sattler, U., Drummond, N., Horridge, M., and Rector, A. (2007). Using owl to model biological knowledge. *International Journal of Human-Computer Studies*, 65(7):583–594.

Suthram, S., Dudley, J. T., Chiang, A. P., Chen, R., Hastie, T. J., and Butte, A. J. (2010). Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. *PLoS computational biology*, 6(2):e1000662.

Swedberg, K., Cleland, J., Dargie, H., Drexler, H., Follath, F., Komajda, M., Smiseth, O., and Tavazzi, L. (2005). The task force for the diagnosis and treatment of chronic heart failure of the european society of cardiology: Guidelines

for the diagnosis and treatment of chronic heart failure: Reply. *European heart journal*, 26(22):2473–2474.

Swinney, D. C. and Anthony, J. (2011). How were new medicines discovered? *Nature reviews Drug discovery*, 10(7):507–519.

Tanimoto, T. T. (1958). elementary mathematical theory of classification and prediction.

Teo, S. K., Stirling, D. I., and Zeldis, J. B. (2005). Thalidomide as a novel therapeutic agent: new uses for an old product. *Drug discovery today*, 10(2):107–114.

ter Horst, H. J. (2005). Completeness, decidability and complexity of entailment for rdf schema and a semantic extension involving the owl vocabulary. *Web Semantics: Science, Services and Agents on the World Wide Web*, 3(2):79–115.

Therapontos, C., Erskine, L., Gardner, E. R., Figg, W. D., and Vargesson, N. (2009). Thalidomide induces limb defects by preventing angiogenic outgrowth during early limb formation. *Proceedings of the National Academy of Sciences*, 106(21):8573–8578.

Tirmizi, S. H., Aitken, S., Moreira, D. A., Mungall, C., Sequeda, J., Shah, N. H., and Miranker, D. P. (2011). Mapping between the obo and owl ontology languages. *Journal of biomedical semantics*, 2(Suppl 1):S3.

Todeschini, R. and Consonni, V. (2009). *Molecular descriptors for chemoinformatics*. John Wiley & Sons.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, pages 433–460.

Van Nueten, J., Van Beek, J., and Janssen, P. (1978). Effect of flunarizine on calcium-induced responses of peripheral vascular smooth muscle. *Archives internationales de pharmacodynamie et de therapie*, 232(1):42–52.

Vempati, U. D., Przydzial, M. J., Chung, C., Abeyruwan, S., Mir, A., Sakurai, K., Visser, U., Lemmon, V. P., and Schürer, S. C. (2012). Formalization, annotation and analysis of diverse drug and probe screening assay datasets using the bioassay ontology (bao). *PloS one*, 7(11):e49198.

Villanueva-Rosales, N. and Dumontier, M. (2008). yowl: An ontology-driven knowledge base for yeast biologists. *Journal of biomedical informatics*, 41(5):779–789.

von Linné, C. and Lange, J. J. (1770). *Systema Naturae: Per Regna Tria Naturae, Secundum Classes, Ordines, Genera, Species, Cum Characteribus, Differentiis, Synonymis, Locis*, volume 3. Curt.

W3C (January 2014a). Class disjointness. `http://www.w3.org/TR/2009/WD-owl2-primer-20090421/#Class_Disjointness`.

W3C (January 2014b). Computational properties of owl2 profiles. `http://www.w3.org/TR/owl2-profiles/#Computational_Properties`.

W3C (January 2014c). Owl 2 web ontology language. `http://www.w3.org/TR/owl2-overview/`.

Wang, H.-Y., Lee, D. H., Davis, C. B., and Shank, R. P. (2000). Amyloid peptide aβ1-42 binds selectively and with picomolar affinity to α7 nicotinic acetylcholine receptors. *Journal of neurochemistry*, 75(3):1155–1161.

Wei, G., Twomey, D., Lamb, J., Schlis, K., Agarwal, J., Stam, R. W., Opferman, J. T., Sallan, S. E., den Boer, M. L., Pieters, R., et al. (2006). Gene expression-based chemical genomics identifies rapamycin as a modulator of mcl1 and glucocorticoid resistance. *Cancer cell*, 10(4):331–342.

Wermuth, C. G. (2006). Selective optimization of side activities: the sosa approach. *Drug discovery today*, 11(3):160–164.

Whirl-Carrillo, M., McDonagh, E., Hebert, J., Gong, L., Sangkuhl, K., Thorn, C., Altman, R., and Klein, T. E. (2012). Pharmacogenomics knowledge for personalized medicine. *Clinical Pharmacology & Therapeutics*, 92(4):414–417.

Wikipedia (January 2014a). Chemical similarity. `http://en.wikipedia.org/wiki/Chemical_similarity`.

Wikipedia (January 2014b). Conservation of mass. `http://en.wikipedia.org/wiki/Conservation_of_mass`.

Wikipedia (January 2014c). Off-label use. http://en.wikipedia.org/wiki/Off-label_use.

Wikipedia (January 2014d). Ontology. http://en.wikipedia.org/wiki/Ontology.

Wikipedia (January 2014e). Paul ehrlich. http://en.wikipedia.org/wiki/Paul_Ehrlich#Magic_bullet.

Wikipedia (January 2014f). Phenotype. http://en.wikipedia.org/wiki/Phenotype.

Wikipedia (January 2014g). Prehistoric medicine. http://en.wikipedia.org/wiki/Prehistoric_medicine.

Wikipedia (January 2014h). Rube goldberg machine. http://en.wikipedia.org/wiki/Rube_Goldberg_machine.

Wikipedia (January 2014i). Science. http://en.wikipedia.org/wiki/Science.

Wikipedia (January 2014j). Thermodynamics. http://en.wikipedia.org/wiki/Thermodynamics.

Williams, A. J., Harland, L., Groth, P., Pettifer, S., Chichester, C., Willighagen, E. L., Evelo, C. T., Blomberg, N., Ecker, G., Goble, C., et al. (2012). Open phacts: semantic interoperability for drug discovery. *Drug discovery today*, 17(21):1188–1198.

Wolozin, B. (2004). Cholesterol, statins and dementia. *Current opinion in lipidology*, 15(6):667–672.

Yang, L. and Agarwal, P. (2011). Systematic drug repositioning based on clinical side-effects. *PloS one*, 6(12):e28025.

Young, A. H. (2011). More good news about the magic ion: lithium may prevent dementia. *The British Journal of Psychiatry*, 198(5):336–337.

Zahler, S., Tietze, S., Totzke, F., Kubbutat, M., Meijer, L., Vollmar, A. M., and Apostolakis, J. (2007). Inverse in silico screening for identification of kinase inhibitor targets. *Chemistry & biology*, 14(11):1207–1214.

Zdobnov, E. M. and Apweiler, R. (2001). Interproscan–an integration platform for the signature-recognition methods in interpro. *Bioinformatics*, 17(9):847–848.

APPENDIX A

Glossary

- **Black box machine**: complex machine composed of a large number of physical parts carrying a certain number of internal functions and acting together in an organised fashion.
- **Description logics**: family of formal languages used to represent the knowledge of a domain of interest.
- **Disease**: impairment of a normal biological condition.
- **Drug**: large or small molecule producing a pharmacological effect when administered in an organism. In this document, this word is interchangeably used with the word *compound*.
- **Drug repositioning**: identification of new therapeutic indications for known drugs. Also referred as *drug repurposing*, *re-profiling*, *therapeutic switching* and *drug re-tasking*.
- **OWL2 EL**: profile of the Web Ontology Language implementing the axioms described in the description logic $\mathcal{EL}^{++}$ family.
- **Gene expression experiment**: quantitative or qualitative characterisation of the number of messenger RNA molecules present in a biological system at a given time and location. This number reflects the expression or activity of genes of interest.
- **Indication**: therapeutic usage of a drug, regulated by an authority.

- **Knowledge base**: set of description logics axioms, entities and expressions.
- **Machine**: assembly of parts functioning to meet a particular goal.
- **Mechanism of action**: biochemical interaction that gives rise to the pharmacological effect of a drug.
- **Mode of action**: broad activity of the molecule on an organism.
- **Ontology**: specification of conceptualisation, formal representation of the knowledge of a domain of interest.
- **Off-label use**: prescription of an approved drug for a non-regulated indication.
- **Off-target**: said to the secondary molecular entities affected by a drug. Off-targets are generally not desirable as they can produce side-effects.
- **Organism**: assembly of molecules functioning as a stable whole.
- **Phenotype**: set of characteristics or traits attributed to an organism.
- **Polypharmacology**: potential of a drug to produce multiple pharmacological effects.
- **Reasoning**: classification and consistency checking of a knowledge base. Also referred as *classification*.
- **Semantic similarity**: value reflecting the closeness between two concepts or entities.
- **Similar property principle**: molecules with similar structures produce similar pharmacological effects.

APPENDIX B

Abbreviations

- **CMap**: Connectivity Map.
- **DLs**: Description logics.
- **GO**: Gene Ontology.
- **GWAS**: Genome wide association study.
- **MoA**: Mode of action.
- **RDF**: Resource Description Framework.
- **SNP**: Single nucleotide polymorphism.
- **OBO**: Open biological and biomedical ontologies.
- **OWL**: Web Ontology Language.

APPENDIX C

Peer-reviewed publications

Chronological list of the peer-reviewed articles published in the course of the thesis. Authors are ranked as present on the original publications. Corresponding authors are underlined. Articles I directly published are openly accessible and a link to the PDF version is provided for the curious reader to follow. Articles resulting from collaborations and behind a pay-wall are identified as such.

- **The CALBC RDF Triple store: retrieval over large literature content.**
  Samuel Croset, Christoph Grabmueller, Chen Li, Silverstras Kavialauskas and Dietrich Rebholz-Schuhmann.
  Proceedings of the Workshop on Semantic Web Applications and Tools for Life Sciences 2010 (2010)
  Document: `http://ceur-ws.org/Vol-698/paper6.pdf`

- **Exploring the Generation and Integration of Publishable Scientific Facts Using the Concept of Nano-publications.**
  Amanda Clare, Samuel Croset, Christoph Grabmueller, Senay Kafkas, Maria Liakata, Anika Oellrich and Dietrich Rebholz-Schuhmann
  Proceedings of the 1st Workshop on Semantic Publishing 2011 (2011)
  Document: `http://ceur-ws.org/Vol-721/paper-02.pdf`

- **OWL Representation of Drug Activities on Biological Systems.**
  Samuel Croset
  Proceedings of the 3rd International Conference on Biomedical Ontology

(2012)
Document: http://kr-med.org/icbofois2012/proceedings/ICBOFOIS2012ECS/SingleFiles/02_ICBO_FOIS_2012_ECS_Croset.pdf

- **Argumentation to Represent and Reason over Biological Systems.**
  Adam Wyner, Luke Riley, Robert Hoehndorf, Samuel Croset
  Lecture Notes in Computer Science - Information Technology in Bio-and Medical Informatics (2012)
  Document (not freely accessible): http://link.springer.com/chapter/10.1007/978-3-642-32395-9_10

- **Integration of the Anatomical Therapeutic Chemical Classification System and DrugBank using OWL and text-mining.**
  Samuel Croset, Robert Hoehndorf, Dietrich Rebholz-Schuhmann
  Proceedings of the Ontologies In Biomedecine And Life Sciences 2012 (2012)
  Document: http://www.onto-med.de/obml/ws2012/obml2012report.pdf#page=23

- **Brain: biomedical knowledge manipulation.**
  Samuel Croset, John P Overington, Dietrich Rebholz-Schuhmann
  Bioinformatics (2013)
  Document: http://bioinformatics.oxfordjournals.org/content/29/9/1238.full.pdf

- **Brain, a library for the OWL2 EL profile.**
  Samuel Croset, John P Overington, Dietrich Rebholz-Schuhmann
  Proceedings of the 10th OWL: Experiences and Directions Workshop (2013)
  Document: http://ceur-ws.org/Vol-1080/owled2013_9.pdf

- **A case study: semantic integration of gene-disease associations for type 2 diabetes mellitus from literature and biomedical data resources.**
  Dietrich Rebholz-Schuhmann, Christoph Grabmller, Silvestras Kavaliauskas, Samuel Croset, Peter Woollard, Rolf Backofen, Wendy Filsell, Dominic Clark
  Drug discovery today (2013)

Document (not freely accessible): http://www.sciencedirect.com/science/article/pii/S1359644613003917

- **The functional therapeutic chemical classification system.**
  Samuel Croset, John P Overington, Dietrich Rebholz-Schuhmann
  Bioinformatics (2013)
  Document: http://bioinformatics.oxfordjournals.org/content/early/2013/11/30/bioinformatics.btt628.full.pdf

- **An ontology for drug-drug interactions.**
  Maria Herrero-Zazo, Janna Hastings, Isabel Segura-Bedmar, Samuel Croset, Paloma Martinez, Christoph Steinbeck
  Proceedings of the Workshop on Semantic Web Applications and Tools for Life Sciences 2013 (2013)
  Document: http://ceur-ws.org/Vol-1114/Session3_Herrero-Zazo.pdf

appendix D

Side projects

List of the finished side-projects directly related to the thesis topic and done within the course of the doctorate. More detailed descriptions of the projects are available online, at the provided URL.

- **PubMed Watcher**: `http://pubmed-watcher.org/`.
- **PubMed**^{2}: `http://pubmed-square.org/`.
- **How cool is my research?**: `http://howcoolismyresear.ch/`.
- **Redesign of the Apache Jena website**: `http://jena.apache.org/`.
- **ChEMBL Twitter Bot**: `https://twitter.com/ChEMBLBot`.
- **XMLBurger**: `https://github.com/loopasam/XMLBurger`.

APPENDIX E

Teaching and editorial work

Summary list of teaching experience and involvement in the scientific peer-review process.

E.1 Teaching experience

- **Network analysis**: introduction to network statistical analysis given to the new PhD candidates entering the EMBL program.

- **Drug-discovery knowledge integration and analysis using OWL and reasoners**: tutorial given at the International Workshop on Semantic Web Applications and Tools for Life Sciences 2012 (SWAT4LS 2012).
  Material: `https://github.com/loopasam/SWAT4LS`

- **Knowledge manipulation using OWL and reasoners for drug-discovery**: tutorial given at the International Conference on Biomedical Ontology 2013 (ICBO 2013).
  Material: `https://github.com/loopasam/owl-tutorial-icbo2013`

- **Supervision of Ashwini Kumar, 6 months Bachelor project internship**: definition of the project, interview of candidates, mentoring and supervision. The topic of the internship was *Exploring RDF at EBI* and resulted in a web application developed by the intern: `https://wwwdev.ebi.ac.uk/chembl/research/reachunique/home`.

E.2 Editorial work

- **Reviewer**: Database (Oxford Journals - `http://database.oxfordjournals.org/`).
- **Program Committee**: International Conference on Biomedical Ontology 2013.
- **Program Committee**: International Conference on Biomedical Ontology 2014.
- **Program Committee**: International Meeting On Computational Intelligence Methods For Bioinformatics And Biostatistics 2014.