

Prediction of biological activity using heterogeneous information sources

PhD thesis

Ádám Arany

Semmelweis University

Doctoral School of Pharmaceutical Sciences



Supervisor: Dr. Péter Mátyus, DSc

Official reviewers: Dr. Gábor Horváth, Ph.D.

Dr. Tóthfalusi László, Ph.D.

Head of the Final Examination Committee: Dr. Imre Klebovich, DSc

Members of the Final Examination Committee: Dr. László Örfi, Ph.D.

Dr. Tamás Paál, CSc

Budapest

2016

1 Table of Contents

2	Abbreviations	4
3	Introduction	7
3.1	Pharmaceutical Industry Background	8
3.2	Overview of Virtual Screening	14
3.3	Overview of Data fusion	17
3.4	Network pharmacology	21
3.5	Evaluating the performance of virtual screening	28
3.6	Probabilistic graphical models in the Bayesian statistical framework	38
3.7	Bayesian framework	38
3.8	Bayesian networks	40
3.9	Bayesian Multilevel Analysis of Relevance	42
3.10	Machine Learning methods	43
3.11	Linear methods for quantitative prediction	43
3.12	Basics of kernel methods	47
3.13	Data fusion with kernel methods	49
3.14	One-class Support Vector Machines	51
3.15	Semi-supervised and Positive and Unlabelled Learning	53
3.16	Generalization error and cross validation	55
3.17	Macau: Bayesian Multi-relational Factorization	56
4	Objectives	60
5	Methods	61
5.1	Information sources	61
5.2	Redundancy and complementarity of the information sources	69

5.3	Evaluation framework for the fusion methods.....	69
5.4	Drug-Indication reference set	70
5.5	Application for Parkinson's disease therapy	71
5.6	Evaluation of Macau	74
5.7	Analysis of the methotrexate pharmacokinetics	76
5.7.1	Patient data	79
5.7.2	Bayesian multilevel relevance analysis	79
6	Results	81
6.1	Fusion of heterogeneous information sources for the prediction of the biological effect of small-molecular drugs	81
6.2	Application of the Kernel Fusion Repositioning method for finding Parkinson's disease related drugs	89
6.3	Predicting multiple activities simultaneously improves the accuracy	97
6.4	Comparison of BN-BMLA results to frequentist statistics in the task of associated variance detection for interpersonal methotrexate pharmacokinetics variability.....	99
7	Discussion.....	102
7.1	Fusion of heterogeneous information sources for the prediction of biological activity	102
7.2	Application of the Kernel Fusion Repositioning framework to find Parkinson's disease related drugs	103
7.3	Prediction of multiple targets simultaneously	104
7.4	Advantages of Bayesian methods	104
8	Conclusions	106
9	Summary.....	108
10	Összefoglalás	109
	References	110

List of own publications118
Acknowledgements119

2 Abbreviations

ADME	Absorption, Distribution, Metabolism, Excretion
ATC	Anatomical Therapeutic Chemical classification system
AUC	Area Under the Curve
BEDROC	Boltzmann-Enhanced Discrimination of ROC
BN-BMLA	Bayesian Network based Bayesian Multilevel Analysis
BPMF	Bayesian Probabilistic Matrix Factorization
CNS	Central Nervous System
COM	Composition of Matter (patent)
CROC	Concentrated ROC
CSEA	Compound Set Enrichment Analysis
DAG	Directed Acyclic Graph
EF	Enrichment Factor
FDA	Food and Drug Administration
FN	False Negative
FP	False Positive
GO	Gene Ontology
GSEA	Gene Set Enrichment Analysis
HCI	High Content Imaging
HDACi	Histone deacetylase inhibitor
HLGT	High Level Group Term (MedDRA)
HLT	High Level Term (MedDRA)
HPO	Human Phenotype Ontology
HTS	High Throughput Screening

IND	Investigational New Drug
INN	International Nonproprietary Name
IR	Infrared (spectroscopy)
ISS	IntraSet Similarity
KFR	Kernel Fusion Repositioning
LLT	Lowest Level Term (MedDRA)
MAF	Minor Allele Frequency
MAO-B	Monoamine Oxidase B
MCMC	Markov-Chain Monte Carlo
MeSH	Medical Subject Headings
MKL	Multiple Kernel Learning
MOU	Method of Use (Patent)
NME	New Molecular Entity
NMR	Nuclear Magnetic Resonance
OLS	Ordinary Least Squares
PCA	Principal Component Analysis
PCR	Principal Component Regression
PGM	Probabilistic Graphical Model
PLS	Partial Least Squares regression
PPAR γ	Peroxisome proliferator-activated receptor gamma
PPI	Protein-Protein Interaction (network)
PT	Preferred Term (MedDRA)
PU	Positive and Unlabelled (learning)
QSAR	Quantitative Structure-Activity Relationship

RBF	Radial Basis Function (kernel)
ROC	Receiver Operating Characteristic curve
ROS	Reactive Oxygen Species
SAR	Structure-Activity Relationship
SEA	Similarity Ensemble Approach
SMARTS	Smiles Arbitrary Target Specification
SNP	Single Nucleotide Polymorphism
SNRI	Serotonin-Norepinephrine Reuptake Inhibitor
SOC	System Organ Class (MedDRA)
SSRI	Selective Serotonin Reuptake Inhibitor
SUI	Stress Urinary Incontinence
TN	True Negative
TP	True Positive
UAS	Universal Average Similarity
UMLS	Unified Medical Language System

3 Introduction

As the productivity of the pharmaceutical research and development is lagging behind the sharply increasing costs, the pharmaceutical industry is continuously searching for new approaches in drug discovery. These problems are aggravated also by the price pressure caused by expiring patents, and the ever complicated regulatory procedures. In my doctoral research I developed and applied methods related to two topics, which revolutionized the pharmaceutical industry to ameliorate the effect of the dropping effectiveness of the research and development pipeline: drug repositioning and personalized medicine (Figure 1).

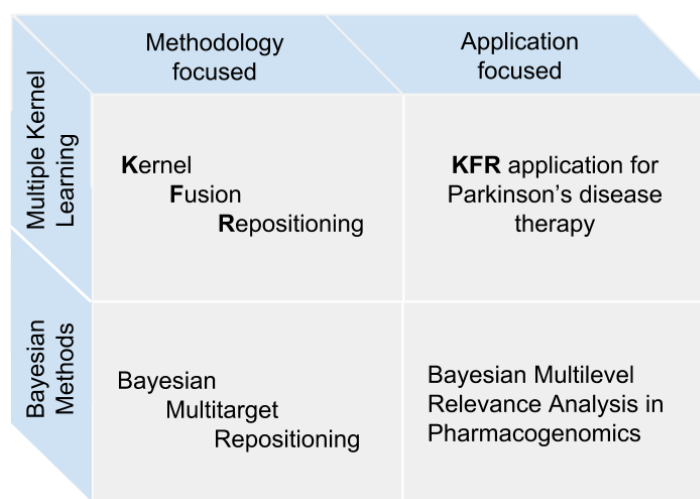


Figure 1 – The investigated topics and their common characteristics.

Drug repositioning or repurposing is a cost-effective and risk-reducing straightforward strategy, which aims at reusing already approved drugs in new therapeutic indications. From the machine learning perspective the main distinctive feature of drug repositioning

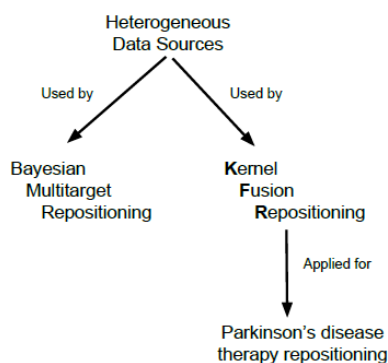


Figure 2 - Topics related to drug repositioning and their relations.

compared to de novo drug discovery is the availability of a wide range of information sources. While conducting the research my primary goal was to develop computational methods to harness these information sources in drug repositioning (see Figure 2). As a first step I created a benchmark dataset containing six different information sources (three chemical structure descriptors, two side effect based descriptors and a target profile), and a drug-indication gold standard set. The goal of my first computational experiment was to compare a novel data fusion methodology, called Kernel Fusion Repositioning (KFR), with a baseline method. My contribution primarily concerned the design and implementation of the KFR framework as well as the application of the KFR framework on the problem of repositioning for Parkinson's disease. As one of the authors of a novel multi-target prediction method I also applied this method to the repositioning benchmark, and analysed the effect of multi-target learning on accuracy.

My second topic was related to personalized medication, which facilitates the optimal therapy for the patient and is also favourable for the researcher interested in drug development. Predicting the patient-by-patient variability of the pharmacokinetics can help the investigator adjust the doses in a personalized way in order to maximize efficacy and minimize side effects and toxicity. I participated in researching the interpersonal variability of methotrexate pharmacokinetics at high dose levels, developed new clinical descriptors bridging patient and treatment levels, and investigated their usage by applying a novel Bayesian multivariate statistical technique to identify predictive genetic variants. Moreover, I compared the results against already existing ones based on frequentist statistics.

3.1 Pharmaceutical Industry Background

The past two decades in the pharmaceutical industry have been characterized by decreasing research and development productivity, high attrition rate and high volatility of output. Nowadays a dramatic shift has taken place in the field, including more pre-competition time collaborations, public-private partnerships, and an extremely high number of mergers and acquisitions [1]. The cost of developing a new drug is steadily increasing, while the yearly number of accepted New Molecular Entities (NMEs) is constant or even decreasing regarding only the small molecular drugs. These trends clearly show that the productivity of the pharmaceutical research and development sector

is deteriorating and the complexity of developing a new drug and the time needed for it is growing significantly [2]. One of the several possible causes of this increased complexity is the stricter regulation environment well illustrated by the increasing number of guidelines [3].

The concentration of research and development efforts in therapeutic areas with larger patient population and higher risk, like chronic and potentially lethal diseases can be observed. A significant exception is the case of rare or orphan diseases, where governmental regulation, like the US Orphan Drug Act and the Regulation (EC) 141/2000 in the EU influence the market [2].

A somewhat radical suggestion to change the patent and the regulatory system has been also discussed in the literature [3]. It is generally accepted that the pharmaceutical industry heavily uses the patent system and employs defensive strategies which can be counterproductive and can further decrease productivity through feedback loops. In the recent years the fear of sharing information seems to ease, but it is still quite prominent. Another way to move the system to a more cooperative mode of operation is to encourage the cooperation between the academia and the industry in a way that protects the academic focus on the long-term goals and high risk innovation.

The history of modern pharmaceutical industry was started by the large scale manufacturing of penicillin. At that time regulation was less strict, only safety studies were required for approval. Essential changes have taken place after the Contergan case, which have led to the introduction of the drug law, the "Arzneimittelgesetz" in Germany, and with a bit logical jump to the Hatch-Waxman act in the USA [4, 5].

In the United States a notable step toward the current regulatory system was the enactment of the Drug Price Competition and Patent Term Restoration Act, often referred to as the Hatch-Waxman Act in 1984. It is useful to shortly examine this law, because it has indirect effects to the pharmaceutical development in the entire world. The main goal of the act is to facilitate generic development and price competition. It is achieved firstly by declaring the sufficiency of bioequivalence studies for generic product approval. A generic company can now enter the market by showing with bioavailability studies that their product is equivalent to the original medicine. The originator always has a 5 year data exclusivity period from approval. During this time the original clinical trials cannot

be used in the registration process of the generic competitor product. As a kind of compensation to stimulate research, patent term restoration has been introduced, which means that half of the time spent between the patent submission and the beginning of the marketing period, but maximum 5 years, can be added to the market exclusivity period of the originator. The whole exclusivity period cannot be longer than 14 years.

In the pharmaceutical industry remarkable volatility of the approval rate appeared in the mid-90s. However a dramatic market entry-exit volatility already existed in the 80s, increased explosively in the 90s and the first decade of the 2000s due to mergers and acquisitions. These trends led to an increasing number of managed NME per organisation. This concentration of patented products led to the birth of a new type of market player, the 'Big Pharma'. Many of these companies do not carry out direct research and development activity or only in limited number, but obtaining NMEs by acquisitions of small companies instead [1].

Historically the time needed for developing a drug from the first screen was 10-15 years, but now this pipeline length is increasing. A target discovery phase, where the main question is the relevance of a target in a particular disease, usually precedes the *de novo* development, but this is out of the scope of the present work. The identification of the compounds starts with *in-vitro* or *in-silico* screening, usually High Throughput Screening (HTS) or virtual screening, where the goal is to search for *hits*, molecules with high probability of target binding. The preclinical development phase in a broader sense includes the classical chemical development steps such as hit-to-lead transition, lead optimization (changing substitution) and synthesis scale up. Strictly speaking the preclinical studies are experiments carried out to prove that the compound is safe to start human studies. These experiments include metabolic stability assays, toxicology studies and limited efficacy studies on model organisms. Before an experimental compound can be tested on human subjects, it need to be registered at the authority as an investigational new drug (IND). The Human Clinical Trials are divided into phases. During the Phase I studies, the main goal is to characterize the safety profile of the compound: determine the maximal safe dose and the absorption, distribution, metabolism, excretion profile (ADME) using increasing doses. These trials are usually carried out with the involvement of 100 or less healthy volunteers. In special indications like cancer a more frequent adverse reaction is acceptable in the hope of the expected favourable risk-benefit ratio. In this case

the safety study, usually called Phase I/II study, involves patients and the determination of a small sample based estimate of the efficacy is also possible. In Phase II the goal is the determination of human therapeutic efficacy with participation of limited number, typically hundreds, of patients. This phase is sometimes divided into sub-phases, like II/A and II/B. The classical setup is a double-blind placebo controlled setting, but it is not always applicable. If an already established therapy exists for the disease, for ethical reasons the control is frequently that existing therapy, and the new compound is given on its own or as an add-on to the classical therapy. The Phase III trial is an extension of the Phase II to 5-8000 patients as a multicentre trial. The successful closure of the Phase III is an essential prerequisite for a regulatory approval. If an already approved drug is tested and found effective in a new indication, a regulatory submission is made for label expansion. This type of submission has considerable interest concerning this work. The process is not finished at the point of approval. The final phase of the compound's lifecycle is the postmarketing phase, or Phase IV, which is about the continuous monitoring of safety also known as pharmacovigilance [6]. The manufacturers together with the medical doctors and the patients continuously monitor adverse events. The continuous data acquisition and interim analysis is pervasive during the whole pipeline both for ethical and financial reasons [7]. In the European Union the 'Community code relating to medicinal products for human use' (Directive 2001/83/EC) outlines the main regulatory background.

Regarding the detailed structure of the unsuccessful cases in the 2011-2012 interval, the main cause of failure in Phase II and III Clinical Trials was the lack of efficacy (56%), followed by safety issues (28%) [8]. The most expensive failure is which happens during Phase III; therefore early termination of the probably unsuccessful projects is an interest of the company. This fail-fast approach can be an explanation for the increase of attrition rate in Phase II and the decrease of failure rate in Phase III. It is worth pointing out that the rate of safety failures increased significantly during Phase III, which can serve as a motivation for this work, because suggesting approved and safe drugs for new indications can decrease the number of safety failures [8]. An analysis from 2014 suggests that the research and development output of the industry is still not satisfactory [9].

Drug repositioning or repurposing, *i.e.* searching for an innovative therapeutic application for an old drug, is a cost-effective and risk-reducing strategy in pharmaceutical research

and development. For an existing compound already approved as a drug in some indications toxicity and pharmacokinetics parameters like ADME profile are available at least in some dosage and for some routes of administration [10]. The already developed manufacturing process or synthesis scale-up can also lower the costs.

The classical case of drug repositioning is when a late failed candidate repositioned to a new indication. The serendipitous observation which led to the repositioning of thalidomide is a good example for this classical route [4]. During a four year period thalidomide with the trade name Contergan was originally marketed as a sedative especially for pregnant women. After its withdrawal due to its serious teratogenic side effects a clinical observation led to its application in *erythema nodosum leprosum*. Similarly, the phosphodiesterase-5 inhibitor sildenafil was in clinical phase for angina, but it failed to favourably influence the clinical outcome. However, its side effect later led to its approval against erectile dysfunction with a trade name Viagra [4].

The case of duloxetine is near to what is called a branching development strategy. Eli Lilly and Co. originally developed the compound as a serotonin-norepinephrine reuptake inhibitor (SNRI) antidepressant later marketed in this indication with the trade name Cymbalta. During its development process, based on a mechanistic observation stress urinary incontinence (SUI) as a new indication was suggested [4, 11]. This repositioning was successful, and duloxetine got approved for SUI with the trade name Yentreve.

We can regard drug repositioning as a lifecycle management, which led to the new trend of early repositioning [12]. The available information during the development process has a funnel structure. The information from the early stages is available for a large set of compounds, it is general, and it can be used independently of the indication, but it is a weak predictor of the clinical outcome. As we proceed, the gathered information will be closer to the clinical endpoint, but its specificity for indications will be higher and higher [12].

In fact the most successful repositioned drugs are based on serendipity, despite the several existing systematic approaches [13, 14]. The evidences show that there is need for technological intellectual property beside of expert knowledge for a repositioning biotech to be successful [13]. An important aspect to understand the difficulty of drug repositioning in a classical pharma company is the management mentality against funding

already failed projects. Another difficulty is that the strategic focus indications are specific for a given company, so if we reposition a proprietary molecule it is highly probable that there is no clinical expertise available in this new indication [13].

Two important types of patent need to be discussed here [4, 13]. The strongest one in the sense of protection is the composition of matter (COM) patent, which claims the chemical structure of the compound and grants 20 years of protection from the patent application. A company needs to protect the compound in development at least before registering it as an IND, so at least half of the time spent in the clinical phases is lost from the market exclusivity period. Therefore, starting a new clinical development phase after a late failure is a risky decision. If a compound fails in Phase III, the company loses too much time to start a new trial: a favourable alternative can be a branched development program [12, 13]. This extended profiling or early repositioning of a drug candidate can facilitate the deeper understanding of the safety and side effect profile of the compound [12, 13].

The other important type of patent in the field of drug repositioning is the method of use (MOU), which claims that the compound can be used to treat a disease. Because the original COM patent usually covers a lot of indications, constructing a MOU patent can be very difficult. Another way to get a new COM patent is the combination of active substances, which forms the base strategy for some of the repositioning biotech companies [4]. In case of the orphan diseases, as already mentioned, an extra protection is granted by the law [13].

Another route to improve the productivity of the pharmaceutical pipeline is the stratification of the patient population. In a more homogeneous population, where a well-defined disease state is present, the lack of efficacy type failures can be reduced [15]. In several cases diseases known in the past as a uniform group actually have several different aetiologies. An excellent example for this is the case of targeted tumour therapies, but the disease heterogeneity is observable in several other therapeutic areas like neurodegenerative diseases or immunological conditions as well [15-17]. This effect can result in apparent inefficacies in clinical trials, as we try to target an ill-defined disease instead of the aetiology.

Drug development for rare or orphan diseases faces with the same complication as stratified patient population: the number of patients can be very low. If we can identify a

common molecular mechanism between diseases, we develop a drug to that mechanism. Another route is drug repositioning: if we can find a drug already registered in a classical indication which can be applied in the rare case, we can use it as a candidate.

Another important factor is the pharmacokinetics related heterogeneity of the subjects. Genetic polymorphisms in metabolizing enzymes and transporters can result in significant differences of drug metabolism and therefore can cause a lack of efficacy and toxicity problems.

3.2 Overview of Virtual Screening

To find novel pharmaceutically active compounds with appropriate properties and a patentable new structure sometimes millions of candidates should be analysed. If we can reduce the number of compounds we need to test in an HTS setting, we can reduce the cost of the early phase of the screening program dramatically. Moreover, sometimes the in-house dataset does not provide enough chemical diversity, therefore we plan to use candidates from external sources or to synthesize new chemistry. In this case, identifying a subset of compounds with the highest probability of activity before candidate acquisition or synthesis would result in even higher benefits.

These computational screening methods can be divided into two main classes, target-based methods and ligand-based methods. In the first case, when the structure of the target is known, this information together with the possibly available structure of known target-ligand complexes can be exploited to guide the search for new active compounds. Most often these target structures are available in the form of X-ray crystallography or NMR measurements. In the other case, ligand-based methods only use the structural information of known active and known inactive compounds and attempts to identify the key elements of the structure-activity relationship (SAR) using statistical techniques. In this work we are particularly interested in ligand-based techniques. The most important categories of these methods are similarity searching, classification and quantitative structure-activity relationship (QSAR) modelling. One of the obvious differences of these methods is that they provide ordinal, categorical and numerical predictions respectively.

In its simplest form similarity searching is a basic tool requiring only a single reference compound, and returning a list of neighbours from the database, ordered from the most

similar to the least similar one. Assuming that the similar property principle holds – two compounds with high global similarity have high probability to share the same biological activity – the biologically active compounds will be enriched on the top of the ranking. The similar property principle is based on the assumption of a smooth structure-activity relationship in the chemical space [18, 19]. While pharmacophore analysis and QSAR based methods focus on local features of the chemical compound, the similar property principle suggests an inherently global viewpoint [18]. This global similarity viewpoint assumes a continuous relation between chemical structure and activity: small changes in molecular structure cause small changes in activity. Therefore its validity is limited by activity cliffs, which can be caused for example by rigid structural elements in the binding site of the target.

A good example for this sudden change of activity in the chemical space is the so called „magic methyl effect”, where introducing a single methyl group to a compound can result in several fold changes in activity. It is hardly surprising that using purely statistical approaches may lead to the misclassification of some samples near to an activity cliff as outlier. Without background knowledge, a sudden change of the activity caused by an activity cliff and a measurement error cannot be distinguished. In spite of the steric limits and well defined pharmacophoric interactions, the structural plasticity of the binding site makes it possible that in practice the biological activity is a smooth function of the chemical similarity in some regions of the chemical space. These factors also explain the strong dependence of the predictive performance on the target protein and on the reference compounds in question [18, 20]. Moreover, the performance depends on the molecular description method used, and on the different binding modes of similar ligands. The above mentioned properties of the chemical space confirm the necessity of data fusion to exploit the advantages of the different methods and reference structures.

To define similarities between compounds we need an appropriate mathematical representation of a molecular structure, on which we can apply a function defining the similarity metric we want to use. The most straightforward approach to represent a chemical compound which meets the requirements described above is to assign a vector of numbers to it. This vector is usually called molecular descriptor. Every position in the vector - either binary, categorical or continuous – encodes a feature of the compound. If the position encodes the occurrence of a substructure, the descriptor is also called

molecular fingerprint. The substructure encoded can be two dimensional or three dimensional. In theory 3D fingerprints would contain more information than 2D ones, but in practice experience shows that most of the time the models based on the former show better predictive performance. A possible cause of this is the uncertainty of the relevant conformation used for calculating the 3D descriptor. Since the 2D fingerprint can be calculated directly from the graph structure of the compound, it is more robust.

Most of the time the number of possible substructures is enormous, while the vast majority of them is missing from a given compound. To handle this situation a function with low collision probability – the hash function - is used to map all possible substructures to a lower dimensional vector. Another solution is folding, when positions in a vector are merged and the new position is set to be active if any of its ancestor was active.

Another problem is that similarity is subjective; as *Maggiore et al.* said „similarity like beauty is more or less in the eye of the beholder” [20]. Or as a machine learning practitioner would say, the selection of the similarity metric should depend on the goal we would like to reach with modelling; that is, it should be determined in a supervised way.

There is a significant difference between classical similarity searching and machine learning methods. This difference is the weighting. Computing similarities between actives and candidates and then applying a predefined similarity threshold does not work, as the optimal threshold depends on the reference compound [20]. When machine learning approaches are used the similarity that we compute will depend on those substructures which are relevant for the binding process, and the possible many irrelevant common substructures will have a low weight.

The most popular similarity metric used on binary fingerprints is the Tanimoto or Jaccard metric. Its most popular chemoinformatics definition in its vectorial form is the following:

$$S(x, y) = \frac{x^\top y}{x^\top x + y^\top y + x^\top y},$$

or written with sets (Jaccard definition):

$$S(X, Y) = \frac{X \cap Y}{X \cup Y}.$$

The Tanimoto similarity is only used on binary vectors in this work. A possible generalization of the Jaccard similarity to non-binary vectors exists for multisets, sets where the number of occurrence of a substructure is also taken into account:

$$S(x, y) = \frac{\sum_i \min(x_i, y_i)}{\sum_i \max(x_i, y_i)}.$$

There are several application areas for molecular similarities, probably the most famous ones are the already mentioned database searching and activity prediction. Molecular similarity also has its application in assessing intellectual property positions and diversity based library enrichment [20]. In these two latter applications our goal is to maximize dissimilarity.

3.3 Overview of Data fusion

Molecules have many different types of measurable or computable characteristics from as simple ones as elemental composition, 2D structure, 3D structure, and physicochemical properties to as complex ones as phenotypic effects in a biological system, which makes the available data very heterogeneous. This heterogeneity is especially high in case of drug repositioning where we can work with much better characterized compounds. The relative importance of these characteristics depends on the scientific question we want to answer. The combination of this type of heterogeneous data should be problem specific, which imposes a significant mathematical challenge. Even in the prediction of drug action, different types of features and different inter-molecular similarity metrics can be predictive for different targets. This type of „no free lunch” characteristic, which is inherent in nature, makes the task even more challenging. The no free lunch behaviour is well known, and mathematically proven in the case of machine learning models [21].

What we call data fusion has always been organic chemists' general practice in a smaller scale. Let us consider a structure elucidation process based on infrared (IR) spectra, mass spectrometry and different multidimensional nuclear magnetic resonance (NMR) experiments. These spectroscopic measurements provide information about the different

aspects of the unknown compound. IR informs us about functional groups with characteristic bands, mass spectrometry about exact molecular mass, and optionally fragmentation data. On the other hand, NMR provides a wide range of information from local environments of protons, carbon, nitrogen, oxygen, phosphorus, fluorine atoms among others, and pairwise bond count distances, or even distances in the three dimensional space. All of these fragments of information, even if some of them are highly redundant, can be used to deduce the structure of the unknown compound with high confidence. In the modern era of big data our need to synthesize information and knowledge is still present, we just need computers to understand the relations in this enormous volume of data.

In chemoinformatics, it is very popular to fuse rankings or scores derived by different methods based on different information sources. In case of the rank fusion, we order the possible candidates (here compounds) and fuse this ranked list to a consensus ranking. This process can be interpreted as a special case of quantile normalization, a statistical technique often used in expression microarray data analysis [22, 23].

In an earlier chemical application of data fusion basic min-rank, max-rank or sum-rank rules were evaluated on the individual rankings [24]. These rules simply calculate the minimum, maximum or average of the given compound's rank in the different lists, and reorder them using this new derived score. The sum-rank rule is commonly referred to as Borda protocol in information retrieval, which name is used in this work. It is originally an election method named after the French mathematician Jean-Charles de Borda. When applying this method, each voter creates a full preference list of the candidates, and scores them inversely to their preference: gives N point to the first candidate, N-1 to the second, and so on. Finally, these scores are summed globally, and the candidate list is ordered based on these points.

If we assume that every scoring function uses only a single reference structure, we can identify two cases. Two scoring functions can be different because the underlying similarity metric used is different, or because the reference active compound is different. The data fusion applied in the former case is called *similarity fusion*, while in the latter case we can talk about *group fusion* [25, 26]. It is shown by multiple studies that in case of similarity fusion sum-rank outperforms min-rank and max-rank, and the average of the

individual data source performance [24]. Furthermore, in some of the experiments the fused score showed at least as good performance on average as the best individual scoring function [24-26].

In case of group fusion it is shown that max-score fusion is better than sum-score or sum-rank [26-28]. Because in this case the underlying similarity metric is unchanged, there is no need for the quantile normalization effect of the rank based fusion rules [27]. As the naming can be misleading at the first read, at this point it should be noted that min-rank is the quantile normalized pair of max-score and max-rank is the pair of min-score. It is also shown that to gain from the application of group-fusion query diversity is preferred. However, it is true that lower query diversity results in higher predictive performance both in the case of the single data source and the fused result. An interesting connection is that one-class support vector machines (see details in Section 3.14) can be interpreted as a robust hybrid of max-score and weighted sum-score rules, where representatives – the so called support vectors – are automatically selected from the reference set. These support vectors represent the boundary of the known actives in the chemical space. In our experiments we found that query diversity is preferred only to a limit (see Section 6.1). The group fusion can be used for scaffold hopping if the fusion strategy is chosen correctly. The simplest max-score rule is expected to result in poor retrieval of new scaffolds, because all of the retrieved molecules will have a single dominant reference molecule, where the similarity is maximal.

A different type of fusion rule beside the rank and score based rules is the voting fusion rule, also called classifier fusion. In this case binary pass-fail votes are aggregated to reach consensus prediction. Votes are collected for all candidates and only those candidates are selected as active which reached a predefined number of pass votes. This fusion technique leads to an increase in precision but a comparable decrease in recall [25, 29, 30].

Unfortunately, it is not possible either to identify the best fusion rule and scoring function combination independently from the target, but fused scores are usually more robust to the change in the task or database than single ones [18, 24, 25]. This shows the real persistent nature of the no free lunch property, and motivates the application of problem specific fusion rules. A possible solution is the application of a regression-based fusion

rule, where the weighting of the different data sources are tuned to get optimal performance on the specific task at hand [31].

Another important decision is how to choose the performance reference for our fusion method. We can compare the fusion result to the best individual data source; in this case the result is clear if the fusion provides a better result than the reference. A less strict reference commonly applied is the average performance of the sources. If our fusion result is better than the average, it is still useful because to select a better than average single data source we need to validate all sources individually, and we will lose statistical power due to the needed validation datasets.

To increase predictive performance in the case of single reference structure *Hert et al.* introduced a method called turbo similarity searching. They used the nearest neighbours of the reference structure as co-reference structures and reached performance improvement [32]. They gave a somewhat ad-hoc interpretation of the result in the paper, but a quite plausible interpretation of the performance gain is a more general statistical phenomenon. The method introduces the local structure of the chemical space into the decision process. In that sense the method can be interpreted as a type of machine learning method from the positive and unlabelled learning class, and shows strong similarities to self-training [33]. For a detailed discussion on the positive and unlabelled learning problem see Section 3.15.

A key assumption in similarity and group fusion is that inactive compounds are more diverse than active ones. This assumption makes it highly probable that different metrics score active compounds more consistently than the inactive ones. We can formulate the independent-and-accurate criteria well known in machine learning as follows: in an ideal case we want information sources which produce accurate ranking on active compounds and uncorrelated ranking on inactive ones. Another formulation for this criterion in the special case of ranking fusion was suggested based on the difference of rank–score graphs [34].

The literature of target-based methods usually refers to data fusion methods as consensus scoring. Ligand poses are evaluated with multiple scoring functions, and then a consensus result is computed [29]. Several fusion rules have been applied for consensus scoring including sum-rank and min-rank like rules, and robustified versions of them where the

worst ranks are dropped before applying the fusion rule. Other approaches use binary pass-fail votes computed based on the different scoring functions, or build regression models to combine scores. Another interesting direction is the combination of ligand-based and target-based data sources [35].

3.4 Network pharmacology

Polypharmacology, a property of a compound to be active on more than a single biological target, has been regarded as unfavourable by the classical medicinal chemistry. Efforts have been made to develop maximally selective compounds, ideally showing high affinity only to a single target. This is a rational approach to reduce the chance of side effects related to off-targets. Paul Ehlich's concept of „magic bullet”, selectively targeting disease causing targets, shaped the landscape of drug design for decades. As the network view of complex diseases got widely accepted, the view of pharmacotherapy as perturbation of a complex network became more and more dominant [36, 37]. Nowadays, the reductionist approach of treating targets as entities standing without biological context is more and more criticized. Psychiatric drugs are typical agents with extensive polypharmacology on central nervous system (CNS) related targets. For example, atypical antipsychotics have activity on a wide range of targets including antagonism on various dopamine and serotonin receptors. Beside the experimental evidences that inhibition of dopamine action on the D₂ receptor seems to be essential for their therapeutic value against the positive symptoms of schizophrenia, other targets - especially 5-HT_{2A} - are also important [38]. Actions on these targets determine the differential behaviour of these agents, like the action against negative symptoms or the risk of dyskinesia. This network view can result in a wider range of information sources for *in silico* methods, including side-effects, off-label uses, molecular biological information and gene expression (see Table 1).

Table 1 - Network levels relevant in the pharmaceutical sciences.

Network Level	Possible information sources
Disease – Disease	Side effect profile, co-morbidity profile
Compound – Protein	Target profile, metabolizing enzyme profile
Protein – Protein	Pathway analysis, target identification
Gene expression	Differential expression profiles (e.g.: CMAP)

The classical target based assay is not appropriate for designing agents with polypharmacology. However, phenotypic screening can be an answer to the problem of modern candidate screening, as it starts from the system level state. In these screens compounds are tested on disease models to achieve a desirable change in phenotype. The downside of this approach is that target deconvolution efforts are needed to figure out the precise mechanism of the candidates found with phenotype based screens.

One class of polypharmacology based therapy can rely on the phenomenon of synthetic lethality. Synthetic lethality is a cellular death occurring due to the simultaneous perturbation of two or more genes or gene products [36, 39]. These perturbations can be caused by genetic change or modification like naturally occurring mutation, knock-out or RNA interference experiment; pharmacological modulation, or environmental changes. Synthetic lethality can be a particularly important mechanism in cancer therapies, where the difference of the tumour cells and the wild-type host cells are in principle characterizable by specific mutations resulting in a changed protein-protein interaction (PPI) network. This new network can have new lethal targets which are non-essentials in the wild-type cells. This approach can be interesting especially in cases where the causal mutation is a loss of function mutation which is complicated to reverse, or it is found in a gene, whose product is difficult to modulate pharmacologically. Similarly, in case of drug combinations where more than one chemical perturbations are applied, the

prediction of the resulting effect needs to take into account the network structure. The detection of these types of complex interactions demands network based multivariate statistical techniques, which can take into account redundancies and synergies between variables. The Bayesian network based Bayesian multilevel analysis of relevance (BN-BMLA) methodology is an ideal candidate for this task (see Section 3.9).

Designing agents for specific disease cases with known genetic variants also leads us to the field of personalized medicine. As in the case of tumour cells, interpersonal variability of the protein-protein network can lead to differences in the set of relevant targets. Therefore, the knowledge of the patient specific network can help choose a therapy which will probably be effective in the case in question counter to the classical therapy effective in the general population.

Synthetic lethality highlights one of the probable reasons why we need compounds with polypharmacology: the well-known robustness of the biological systems. As developed by evolutionary steps under continuously changing environmental conditions, these complex systems need to be robust against most of the single point changes and against a wide range of environmental effects. We need network biology based considerations to attain stable changes of the phenotype [37, 40].

Modulating central protein nodes, hubs, with a really high number of connections, can lead to toxicity because of the essentiality of these proteins. Conversely, peripheral nodes are probably well buffered, and drugs acting on these targets can have a lack of efficacy type problems. It is found that the middle ground, highly connected but not essential proteins are good drug targets. According to the network pharmacology paradigm the goal is to identify one or more network nodes – target candidates – whose perturbation would result in system level changes, and, more importantly, a favourable change in the disease related phenotype.

An interesting new direction is the intentional design of multi-target directed ligands, using the already known SARs [16]. One possible option is the design of conjugated ligands when two or more already tested bioactive pharmacophores are linked together to form a new ligand. This method can result in high molecular weight and ADME problems. Another technique is to design a ligand with overlapping pharmacophores which can lead

to smaller molecular weight and structural complexity, but at the same time makes the design process more complicated.

The method of selective optimization of side activities (SOSA) can also be used as a route to polypharmacology. The main idea of SOSA is to screen a diverse set of existing drugs for new activities with the aim of finding a starting point for further optimization, and not a candidate for direct repositioning [41]. With this method all starting points will be drug like by definition. The optimization goal thereafter will be twofold: on the one hand, to increase the new activity of the candidate; and on the other hand, to reduce the old activity. In case of optimization for polypharmacology, the original activity can be one of the desirable activities.

Screening methods using gene expression become a universal reductionist approach. The proposal of gene expression as *lingua franca* of different perturbations on a biological system had a great impact [42]. The Connectivity Map defines a biological state by a gene expression profile, which is clearly a reductionist approach given that the downstream state variables like protein and metabolite levels and post translational modifications are not included.

The Connectivity Map contains a database of reference profiles; gene sets ordered by differential expressions in a control–treatment setting. Using a query signature, a list of differentially expressed genes annotated by the direction of the expression change, the reference databased can be searched. The retrieved profiles are then ordered based on a gene set enrichment score, called connectivity score. The score can be positive or negative depending on the relative direction of the differential expressions. If the directions are the same in the query signature and the database profile the connectivity score is positive, but if they are reverse, the score is negative. The original work suggests that if a perturbation *A* has negative connectivity score with condition *B*, then it may reverse the effect of the condition. In practice this is true only if a strong linearity assumption of gene expression changes holds.

Chemical compounds, short hairpin RNAs or, more generally, perturbagens can be used to treat different cell lines. In the Connectivity Map reference set relatively high concentrations (mostly 10uM) and short accumulation times (mostly 6h) were often used.

This time is usually not enough for feedback loops to get activated, and to cause changes in the expression of the target itself [43].

Illustrative examples on histone deacetylase inhibitors (HDACi), oestrogens, phenothiazines and natural compounds show that the method can recover structurally non-related ligands, can differentiate between agonists and antagonist and can be used for target discovery [44]. The usage of disease related profiles from an animal model was also demonstrated on the case of connectivity between diet-induced obesity profiles and peroxisome proliferator-activated receptor gamma (PPAR γ) inhibitors. Two demonstrative examples were also given for human samples: Alzheimer's disease and dexamethasone resistance in acute lymphoblastic leukaemia [44].

A similar connectivity database was also built from differential expression profiles based on Gene Expression Omnibus DataSets [45]. A network containing disease and drug nodes and edges between them was constructed using profile correlation or using the same signature enrichment based method as in Connectivity Map. The set of nodes in the network was also extended with the reference profiles from the Connectivity Map. It was illustrated that the disease–drug links in this network can be used as hypotheses for drug repositioning and side effect discovery; while on the other hand, drug–drug links can be useful in target and pathway deconvolution.

A network based analysis method for differential expression in these chemical perturbation experiments discussed above is also suggested [43]. This method uses functional protein associations from a database of known and predicted protein-protein associations. It is shown that simple differential expression based ranking is not a good predictor for target identification, because it relies on feedback mechanisms changing the own expression level of the targets. Therefore, a diffusion method is used to distribute differential expression based evidences through the network. These evidences are diffused through the functional association links, or based on the correlation of the neighbourhood structure of the proteins. It is not surprising that this method works best on nuclear receptors, which are directly linked to the gene expression level. Galahad, a free online service based on this method provides full microarray data processing pipeline for drug target identification [46]. It can be used to prioritize candidate targets, predict new mode of actions or off-target effects.

The network view also changes the way how we see diseases. Contrary to the traditional symptom based classification, more and more effort is made to discover the common mechanisms, and the co-morbidity structure of diseases. A good illustration for the entanglement of disease states is the fact that a naive guilt-by-association based method can reach surprising performance [47]. The suggested method is based on the following assumption: if two diseases share a drug, another drug for one of the diseases can be prioritized as treatment of the other. During the evaluation of the method 12 fold enrichment has been detected in clinical trials relative to random drug–indication pairs.

The similarity of the active ligands on two proteins is a more sophisticated information which can be used. The binding site similarity of two proteins can be significantly different from their sequence similarity and can be unrelated from their evolutionary origin. A common endogenous ligand in a metabolic pathway for example can result in a convergent evolution of the binding sites. A similar phenomenon is the existence of ionotropic and metabotropic receptors for the same endogenous ligand. Based on this observation, a method called the Similarity Ensemble Approach (SEA) was developed [48]. SEA assesses protein similarity using 2D fingerprint based similarity of their ligands. More precisely it analyses the distribution of pairwise Tanimoto similarity scores between ligands of the two proteins with a correction for set size bias. Analysing the differences between sequence based and ligand based similarities shows typical protein groups with divergent and convergent binding site evolution, furthermore it illustrates the current trend of selective ligand design. It has been illustrated that the method can be used for the prediction of new primary or side effect related targets even between protein families [49].

An approach with possible application for personalized medicine is also suggested in the literature [50]. This method can handle repositioning scenarios and novel molecules as well. Using known associations as gold standard for training a classifier to distinguish valid associations from random pairs, the method can be seen as a multi-task learning method. Because data are only available for valid drug–disease associations, random associations are used as a negative set for training. For more details on methods for learning from positive and unlabelled samples see Section 3.15. The method uses five drug–drug similarities (three out of which are drug target related) and two disease–disease similarity metrics to describe the associations. The applied drug–drug similarities cover

the chemical and side effect aspects, and the similarities between the drug targets based on sequence, PPI network and Gene Ontology (GO) categories. The disease–disease phenotype similarities are based on Medical Subject Headings (MeSH) terms and Human Phenotype Ontology (HPO) base semantic similarity. An alternative set of disease–disease similarity based on gene expression signatures was also used. This points to the direction of personalized medicine: diseases can be represented with expression profiles, therefore a given specific case of the disease can be screened as well. After the application of a conservative cross-validation scheme the method reached significant predictive performance. A biologically motivated validation technique was also applied based on disease–tissue and drug–tissue associations. The hypothesis behind this validation was that it is highly probable that a target of a drug should be expressed in the tissue, which is relevant in the context of the new indication.

The side effect resource (SIDER) developed by *Kuhn et al.* contains side effect terms and frequencies of occurrences based on text mining from public data sources, mainly FDA package inserts [51]. For text mining a side effect dictionary based on the Unified Medical Language System (UMLS) ontology has been used. As side effects can be regarded as phenotypic responses to a given chemical perturbation, they represent valuable information for describing biologically active compounds. Placebo controlled frequencies have been also extracted for a subset of the drugs.

It is shown by the same research group, that the set of side effects can be used as a predictor for drug–target interaction in the context of drug repositioning [52]. The above discussed work of this group, which was one of the main motivations of the repositioning related works in our research group, led to a patent application about aprepitant as a potential agent in cancer therapies [53]. It is claimed that aprepitant is a non-competitive inhibitor of the enzyme thymidylate synthase and inhibits cell proliferation.

PROMISCUOUS is another online database project; it is a rich information source with search and network exploration tools with the purpose of helping drug-repositioning [54]. PROMISCUOUS contains four different types of interactions; namely, drug–protein, protein–protein, drug–side effects and drug–drug, where protein targets are also mapped to KEGG pathways. There is a possibility to search the database by drug, ATC class, side effects, targets or KEGG pathways, and to visualize the interaction in a network. The

system has a side effect similarity feature, which is able to list drugs based on a high number of shared side effects.

3.5 Evaluating the performance of virtual screening

First, we illustrate performance measures in a medicinal chemistry context using a small example. Let us assume we have 20 unknown compounds, 5 out of which are active COX-1 inhibitor. The fraction of actives (R_p) in this dataset is 25%, which is selected for illustrative purposes and unrealistically high in practice. We have three different methods which order these compounds based on the chemical structure and further information we have. After ordering the compounds, we check the COX-1 inhibitory activity of the compounds *in vitro*, and we get the result on Figure 3.

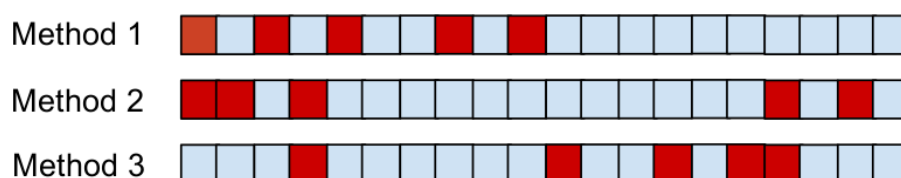


Figure 3 - Output ordering of three hypothetical prioritization methods on 20 compounds: 5 active (red boxes) and 15 inactive compounds (blue boxes). Predicted activity is highest on the left side and lowest on the right side.

In this example red colour always indicates active compound and blue indicates inactive compound. In Figure 3 boxes represent compounds in the order the given method ranked them. The question is which method is the best.

The answer, as always, depends on what we mean by 'the best'. To concentrate on performance measures, let us regard these predictors as black boxes, which means there is no use of the predictor on its own; we want to evaluate only the *predictive performance* of the models. For instance, we cannot learn new chemistry by inspecting them, or we cannot use the models for lead optimization more effectively than just predicting activities of analogues.

If we have limited capacity to test compounds, and we want good candidates for our pipeline, how much can we gain? We will apply a threshold τ to define the position in the

list above which the compounds are predicted to be active. To measure predictive performance we need to define some statistical measures:

True Positive (TP): Number of compounds our classifier predicted as active out of the real active ones.

False Positive (FP): Number of compounds our classifier predicted as active but which actually are inactive.

True Negative (TN): Number of compounds our classifier predicted as inactive out of the real inactive ones.

False Negative (FN): Number of compounds our classifier predicted as inactive, but which actually are active.

All of these four measures are threshold dependent, therefore we could write them as functions as well in the form: $TP(\tau)$, $FP(\tau)$, $TN(\tau)$ and $FN(\tau)$ respectively. We will need two parameters of the library, which are independent of the model and the applied threshold:

N_P (All Positives): The number of active compounds in our library. N_P which applies to our whole library is usually not known, but it can be known in case of a validation set.

N_N (All negatives): The number of inactive compounds in our library. The sum of N_P and N_N is the size of the library N_A , and the ratio of actives can be written as $R_P = N_P/N_A$.

As we will see, these numbers will be sufficient to derive all measures we need in a contingency table (Table 2).

Table 2 - Contingency table. Table containing all sufficient statistics we need to assess performance given a threshold τ .

	Real active (N_P)	Real inactive (N_N)
Predicted to be active	True Positive	False Positive
Predicted to be inactive	False Negative	True Negative

We will use the following derived measures:

Sensitivity (also called Recall): The fraction of the active compounds that the classifier identified successfully. TP / N_P

Specificity: The fraction of the inactive compounds that the classifier excluded successfully. TN / N_N

Precision (also called Positive Predictive Value): The fraction of real actives, in the set of compounds that the classifier identified as active. $TP / (TP + FP)$

For a medicinal chemist, precision has a probably more intuitive form called the *Enrichment Factor* (ER). ER is a normalized form of *precision* by the fraction of active compounds in the whole dataset. It measures the fold of increase in the number of hits, which the experimenter can get if instead of choosing random compounds from the library, they test compounds predicted by the model. We can write EF proportional to sum of weights for all active compounds [55]:

$$EF = \frac{\sum_{i=1}^{N_A} w_i}{\tau R_P},$$

where the weighting for a compound ranked before the threshold is 1, and after the threshold is 0:

$$w_i = \begin{cases} 1, & r_i \leq \tau \\ 0, & r_i > \tau \end{cases}$$

where r_i denotes the rank of the active compound i in the output ordering of the model.

Let us assume that we have capacity to test 4 compounds, so we will use the methods to predict the 4 most likely active compounds (Figure 4).

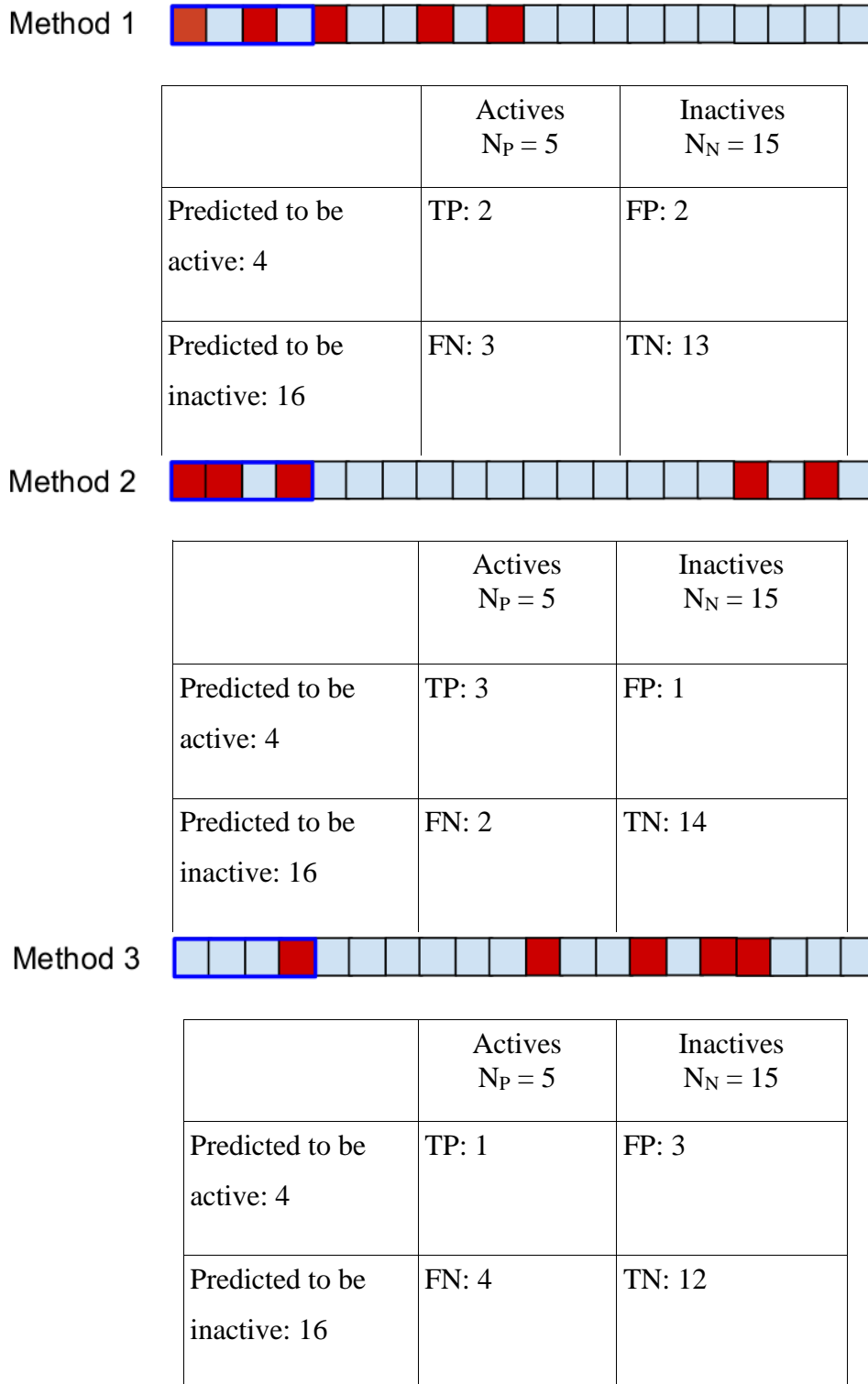


Figure 4 – Contingency tables and graphical illustration of thresholding for $\tau = 4$

Based on these contingency tables, we can compute the derived measures (Table 3).

Table 3 - Derived measured computed for $\tau = 4$

	Method 1	Method 2	Method 3
Sensitivity	0.40	0.60	0.20
Specificity	0.87	0.94	0.80
Precision	0.50	0.75	0.25
Enrichment Factor	2.00	3.00	1.00

For example, in case of Method 1 40% of the actives (two out of five) has been in the 4 selected compounds, therefore they have been identified successfully, while 87% of the inactive compounds are excluded. The ratio of actives in the 4 selected compounds is 50%, which corresponds to a two fold increase relative to random selection. We can see from Table 3 that the *classifier* corresponding to Model 2 at the selected *threshold* outperforms the other two classifiers irrespectively of our optimization goal. For example, this classifier improves our hit rate by 3 folds. While this classifier is objectively better, the same is not true in the level of the methods. Let us choose now a new threshold: we can now test 10 compounds (Figure 5).

Method 1 

	Actives $N_P = 5$	Inactives $N_N = 15$
Predicted to be active: 10	TP: 5	FP: 5
Predicted to be inactive: 10	FN: 0	TN: 10

Method 2 

	Actives $N_P = 5$	Inactives $N_N = 15$
Predicted to be active: 10	TP: 3	FP: 7
Predicted to be inactive: 10	FN: 2	TN: 8

Method 3 

	Actives $N_P = 5$	Inactives $N_N = 15$
Predicted to be active: 10	TP: 1	FP: 9
Predicted to be inactive: 10	FN: 4	TN: 6

Figure 5 - Contingency tables and graphical illustration of thresholding for $\tau = 10$

The computed derived measures are shown in Table 4.

Table 4 - Derived measures computed for $\tau = 10$

	Method 1	Method 2	Method 3
Sensitivity	1.00	0.60	0.20
Specificity	0.67	0.53	0.40
Precision	0.50	0.30	0.10
Enrichment Factor	2.00	1.36	0.40

With this threshold the classifier corresponding to Model 1 outperforms the other two classifiers according to all measures. It is clear that the performance of a model we want to use for classification will depend on the threshold, but will not depend on the ordering of the compound above or below that threshold. We can plot this performance for all possible thresholds using a tool called Receiver Operating Characteristic or ROC curve. As a convention, we plot the *sensitivity* with respect to *1-specificity* for all threshold levels (Figure 6).

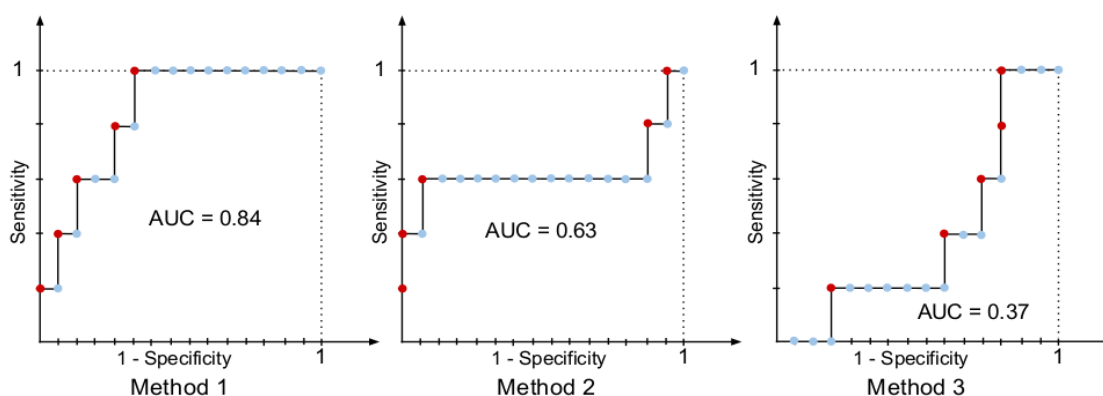


Figure 6 - Receiver Operating Characteristics (ROC) curve of the three different prioritization methods. Every coloured dot corresponds to an active (red) or inactive (blue) compound in the ordered sequence.

In some cases we are interested in the predictive performance of the models in a threshold independent way. One measure to use in this case is the area under the ROC curve (AUC). The AUC value has a very intuitive interpretation: it gives the probability of ranking an active compound higher than an inactive one if the inactive-active pair in question is drawn uniformly at random. This interpretation relies on the connection between AUC and the Mann-Whitney-Wilcoxon statistics [56].

We can see that according to the AUC measure, Method 1 is better than Method 2, which is better than Method 3. We can also realize that Method 3 would be a bit better if we inverted its ordering. The truth is that the ordering for Method 3 was generated randomly; and because of the small number of entities, its AUC value can randomly deviate from the totally random model. If we had a huge number of entities, a random model would be a diagonal line with $AUC = 0.5$. To better understand the apparently controversial statements about Model 1 and Model 2, let us examine the ROC plots of them together (Figure 7, left).

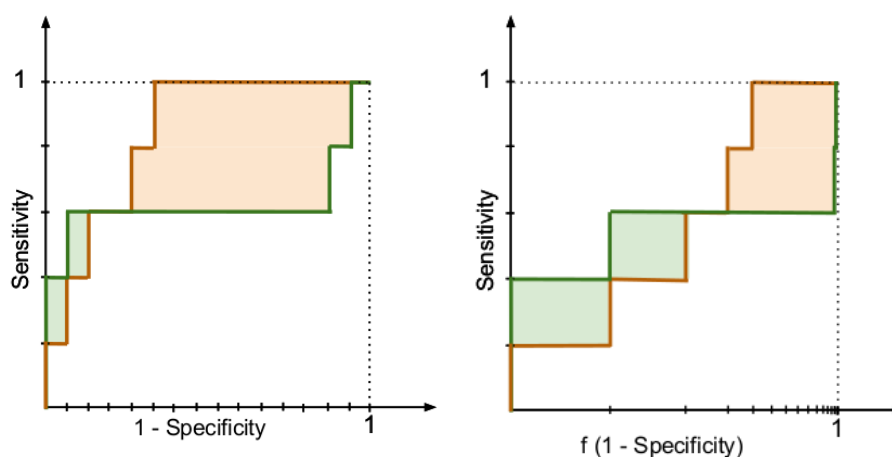


Figure 7 - Comparison of the Receiver Operating Characteristics (ROC) (left) and the Concentrated ROC curves (right) of Method 1 and Method 2. The green area represents the superiority of Method 2 in classification tasks using small threshold value, the brown area represents the superiority of Method 1 in case of high threshold values.

The green ROC curve corresponds to Method 1, and the brown one to Method 2. The green and brown shaded area corresponds to the superiority of Method 1 and 2 respectively. The AUC metric weights the two sub-area equally, Model 1 therefore has higher AUC value.

It can be shown that we can write AUC in the following form:

$$\text{AUC} = \frac{1 - \frac{1}{N_A} \sum_{i=1}^{N_A} \frac{r_i}{N_A}}{1/R_P} + 2R_P,$$

where r_i denotes the rank of the active compound i in the output of the model. From this equation we can see that AUC is a linear transformation of the sum of a weighting, where the weights are:

$$w_i = \frac{r_i}{N_A}.$$

If our chemical library contains several millions of compounds, but we have a limited testing capacity for testing only the top hits – which is the case in practice –, we are usually not interested in the performance after the top hits *i.e.* in the brown area. We want to weight the early part higher, and only invest time and money in more measurement, if it is really worth it. In this case we are facing the so called *early recognition problem*. An

intuitive way to do it is to transform the horizontal axis of the ROC curve in such a way, that the area elements in the early part of the curve will be magnified, and in the late part they will be compressed. In a more formal way, we apply a continuous compression function f to 1 -specificity, which maps the $[0,1]$ interval to itself, $f: [0,1] \rightarrow [0,1]$. An illustration is shown on Figure 7 (right), where we can see that the green area in the low end of the 1 -specificity axis is now magnified, while the brown area at the high end is compressed. This type of transformation reflects our preference in the early recognition problem and can be achieved by a concave compression function, which has a derivative higher than 1 for low values, and lower than 1 for high values. This measure is called the concentrated ROC (CROC) [57]. A well-behaving compression function, which we will use in this work is the exponential compression:

$$f(x) = \frac{1 - e^{-\alpha x}}{1 - e^{-\alpha}}.$$

This function has a parameter α , which defines how early is the part we want to focus on.

A very similar measure can be derived also from the probability theory point of view, called Boltzmann-Enhanced Discrimination of ROC (BEDROC) [55]. If we enforce that our weighting corresponds to a proper probability density function $f(x)$, then we can interpret our weighted sum in the continuous limit:

$$\text{wAUAC} = \int_0^1 \text{TP}(x) f(x) dx$$

as an expected value, which have a similar probabilistic interpretation to AUC given that $\alpha N_P/N_N \ll 1$ holds, which is usually the case in virtual screening, and is the case in our experiments as well.

Similarly to the case of AUC, we need to transform this rank average based metric to get a metric which have values between 0 and 1:

$$\text{BEDROC} = \frac{\text{wAUAC} - \text{wAUAC}_{\min}}{\text{wAUAC}_{\max} - \text{wAUAC}_{\min}}.$$

It is important to note that BEDROC is not a measure which tells if our model is better than a random model or not. It tells if our model is better than a reference enrichment, we

selected that way that it is sufficient to reach our virtual screening goal. We can however, analytically compute the BEDROC score of a model ranking the actives corresponding to a uniform distribution:

$$\text{BEDROC}_{uni} = \frac{1}{\alpha} + \frac{1}{1 - e^{-\alpha}} \quad \text{if } \frac{\alpha N_P}{N_N} \ll 1$$

which is with a good approximation 0.05 for $\alpha = 20.0$.

In case of the sensitivity, precision or EF, a hard threshold is applied, which means every active compound found higher than the threshold level is counted with equal weight, and every compound under this level is ignored – weighted by zero. In case of AUC all active ranks have equal weight. The measures like CROC or BEDROC can be interpreted as a trade-off between these two extremes: a decreasing weighting function - in our case an exponential - is applied to the ranking. These measures can take into account the early discovery requirement, but they are more stable than hard thresholded methods. Because we want to apply our method not exactly on the compound library we used for testing, but possibly many similar libraries, we need a robust evaluation, which does not depend strongly on small perturbations of the order.

3.6 Probabilistic graphical models in the Bayesian statistical framework

Probabilistic Graphical Models (PGMs) are standard representations of complex probabilistic models, as they allow the use of the underlying independencies in both model specification, learning and inference. A particularly popular subclass of PGMs are the Bayesian networks (BNs), which allow the specification of local dependencies.

Another universal framework used in the thesis is the Bayesian statistical framework. In the following sections an overview will be given about the basics of that methodology in the hope that the name will gain its correct semantics.

3.7 Bayesian framework

The Bayesian statistical framework is gaining wider and wider acceptance as a principled approach to cope with uncertainty with respect to *a priori* knowledge, statistical models

and predictions [58]. In the Bayesian framework we do not assume that we will be able to build „the correct model” based on a limited number of observations, therefore we use all possible models weighted by their probability of correctness; the probability distribution of all models. In practice we calculate with a limited set of probable models because of the computational limitations. As the Bayes-theorem states, after which the framework is named:

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)}.$$

Where $P(M|D)$ denotes the above mentioned probability distribution of all models M given the observed data D , called the *posterior* distribution. $P(D|M)$ is the data likelihood, the probability that we observe the data we have, given that M is the correct model. $P(M)$ represents the *a priori* correctness assumption of the models. $P(D)$ is the marginal likelihood. As it does not depend on M it is only a normalization constant of the probability density over the models. From now on, we can use this $P(M|D)$ distribution for prediction, or to gain understanding of the dependency structure between the variables in the domain. We can infer the *a posteriori* distribution of a variable V using marginalization, *i.e.* summing the distributions according to each model, weighted by the model probability:

$$P(V|D) = \sum_M P(V|M)P(M|D).$$

Similarly, we can compute marginals not only to get the distribution of a variable or set of variables in the domain, but to get the distribution of some properties of the model as well. Let $f(M)$ denote a feature of a model, like the existence of some kind of statistical dependence between two variables. Then we can compute its *a posteriori* probability by:

$$P(f(M)|D) = \sum_M f(M)P(M|D).$$

3.8 Bayesian networks

Bayesian networks are graphical models with directed acyclic graph (DAG) structures. They are composed of a set of vertices (representing random variables) and directed edges between them. Being a DAG they do not have directed circles, which statement is equivalent to the fact that at least one ordering of the variables exists in which edges are only directed from variables to other variables forward in the ordering. Those variables

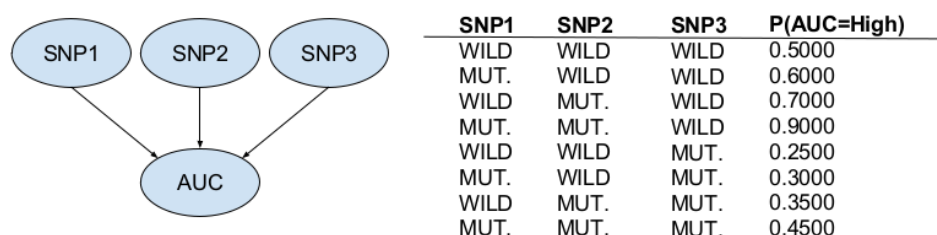


Figure 8 - Local probability structure of a variable and its parents.

from which variable V has incoming edges are called the parents of V . Here we are interested only in the structure of the Bayesian networks; we will ask questions related to the existence of some edges and neighbourhood relations. These relations encode direct relevances of genetic polymorphisms to measured pharmacokinetic parameters of drugs. However, to provide an insight into how the probabilities of the possible models are constructed we should take into account that there is an other layer of specification. Every vertex has a description of the conditional probability of the variable it corresponds to, given the value of its parents. In general, this description can be any function mapping from the values of the parents to a distribution of the child variable. However, in the case of discrete variables a tabular representation is usually used. To illustrate the concept of these local probability models on an example, let us consider the local structure on Figure 8. From the conditional probability table we can read that if all relevant single nucleotide polymorphisms (SNP) are absent, the probability of measuring a high Area Under the time plasma concentration Curve value is 0.5. If the other SNPs are absent, $SNP1$ increases the odds of a high AUC by 50%, $SNP2$ by more than two folds and the two together by 9 folds.

Every variable from which we can go to the variable V by stepping through edges in the direction corresponding to its orientation are said to be an ancestor of V . It can be seen that every variable is independent of these ancestors if we know the value of their direct parents.

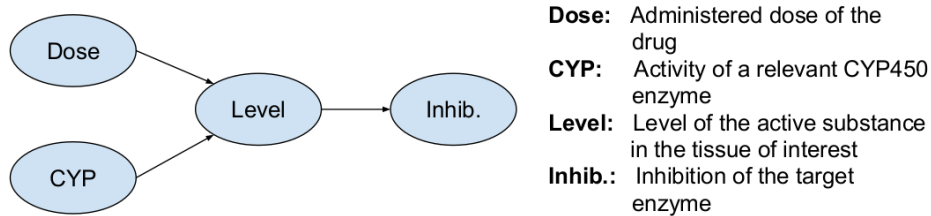


Figure 9 - A simplified illustrative Bayesian network of a drug action. The administered dose (Dose) and the activity of the relevant CYP450 enzymes (CYP) directly influence the level of the active substance (Level) in the tissue of interest, given that no other variable is known. This structure does not state that these are the only influencing variables, neither that there is no other intermediate variable, like the plasma level of the drug in the chain.

In the example of Figure 9 the parent of the *Inhib* variable is only the *Level* variable, which means that according to that model, if we know the active substance concentration in the tissue we are interested in, the *dose* of the drug or the *CYP* enzyme activity gives no more information about the potential inhibition of the target. This is true only in the case of ancestors. However, if we know the value of the *Dose* and *CYP* it is still possible that we can gain extra knowledge about the concentration of the active substance by measuring the inhibition. To rule out or isolate all effects, we need to know the values of all variables in the Markov blanket of our variable. The Markov blanket of a variable contains its parents, its children, and the other parents of its children. Two of these three cases was previously discussed. Let us assume that the mechanism of our hypothetical drug is a competitive antagonist. In that case the inhibition is also influenced by the level of an agonist in the tissue, because the agonist can overcompete the inhibitor in the binding site [59]. In Bayesian network terms we have a new parent, let us call it *Agonist*, of the *Inhib* variable. Now the knowledge of this *Agonist* concentration can be important to infer the distribution of the *Level* variable from the inhibition.

3.9 Bayesian Multilevel Analysis of Relevance

The Bayesian network based Bayesian multilevel analysis of relevance (BN-BMLA) provides an overview of multivariate strong relevance relations, including the option of multiple target variables in a multi-task setting. The BN-BMLA uses hierarchical, systematically linked levels of representations, such as Markov Blanket Memberships (MBMs), Markov Blanket Sets (MBSs), and their subsets (k-MBSs). The method was applied in a wide range of bioinformatics problems including genetic association analysis [60-64].

Using the marginalization formula for model properties, we can determine the probability of some features in the Bayesian network even if we do not have enough data to determine the complete structure.

A simple $f(M)$ function is used to discover edges. The value of this function is 1 if there is directed edge from variable A to variable B , and zero otherwise. Another feature function used in this work is the MBM function. The MBM function is 1 if variable A is in the Markov blanket of variable B and zero otherwise.

There is a possibility to calculate edge or MBM features to a set of target variables. In that case the value of the function is one, if there is an edge from variable A pointing to any variable in the predefined set of target variables, or analogously if A is in the Markov blanket of at least one variable of the target set.

There are several other possibilities, but they are not relevant to this work and they are discussed in detail in a book chapter published by the research group at the Technical University [65].

3.10 Machine Learning methods

In the chemoinformatics literature the main distinctive feature of machine learning methods relative to similarity searching methods is the use of the inactive compound set [25]. Here we assume that the discussed methods are black boxes; we are only interested in predictive performance and not in interpretation. The machine learning community defines itself in a broader sense, e.g. also including methods using only positive labels with the goal of learning their weighting. In this work the latter convention is used, therefore we discuss here one-class methods, and semi-supervised learning methods in the extreme case where only positive samples are available.

3.11 Linear methods for quantitative prediction

In *regression setting* our goal is to build a quantitative model of one or more outcome variables using features also called independent or explanatory variables or covariates. In chemoinformatics the main applications for regression models is the field of quantitative structure-activity relationship modelling (QSAR). For the discussion of this setting, let us assume that the features are organized in an N-by-F matrix X where a row corresponds to a sample – here compound – and a column corresponds to a feature. Furthermore let us organize all outcome variables to an N-by-M matrix Y , where a row corresponds to a sample and a column to an outcome variable. See the illustration on Figure 10. In the following discussion - if otherwise not specified - we will work with a single outcome (*univariate regression*). In this case Y is a column vector.

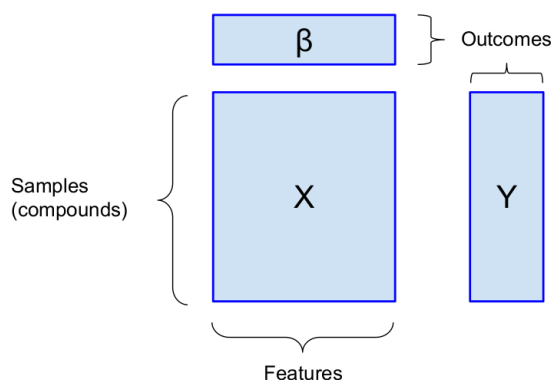


Figure 10 - The structure of a linear regression problem in its general multivariate form: X is a sample by feature matrix containing the samples of the covariates, Y is the outcome matrix, and β is an outcome by feature weight matrix, containing the model parameters. As a convention we add a feature which is always one, and the β corresponding to that feature is the bias of the model.

Ordinary Least Squares (OLS) is the simplest form of regression methods, which can be used to predict compound activities. The name comes from the fact that the method minimizes the squared error between the prediction and the known outcome:

$$\sum_{i=1}^n (y_i - x_i^T \beta)^2,$$

where β is a vector of model parameters, interpreted as weights on the elements of the feature set. The vector x_i is a row of the matrix X corresponding to sample i , and y_i is the value of the outcome variable corresponding to sample i .

It can be shown that the β for which the above error term is minimal can be calculated as:

$$\hat{\beta} = (X^T X)^{-1} X^T y,$$

where the expression multiplying y is called the Moore-Penrose pseudoinverse of X . If two features are linearly dependent – they differ only by a linear transformation plus a small deviation term – exchanging and transforming the two β values will result in similar predictions. From the other way around, a small change in y would result in a huge change in some β values. In mathematical terms the condition number of the matrix will be large. Even if the problem is numerically stable, models with a high dimensional feature set

trained on a small number of training samples can have suboptimal performance. See the topic of over-fitting discussed in chapter 3.16. We can ameliorate these problems if we introduce a constraint to restrict the space of possible models, often called *regularization*. One possible way is to reduce the actual dimensionality of the feature set by Principal Component Analysis (PCA). PCA will find new derived features which are uncorrelated. It can be interpreted as finding a transformation of the coordinate system to minimize correlation between the new variables (see Figure 11). In this case feature 1 and 2 are nearly linearly dependent, which would cause numerical instabilities during the computation of the Moore-Penrose pseudoinverse. On the other hand, the principal component corresponding to the largest variance (PC1) and the second principal component (PC2) are totally independent, and PC2, which corresponds to the deviation from the linear dependence, has small variance. If the two features were perfectly dependent, PC2 would have zero variance. More formally PCA finds two matrices U and V satisfying

$$X = UV^{\top} + E,$$

where U is a sample-by-principal component matrix, called the *score matrix*, and V is a feature-by-principal component matrix called the *loading matrix*. The rows of U , or simply scores, describe the samples in the new space, as plotted on Figure 11. The rows of V , or loadings, however, define the transformation from the original to the new feature space.

The technique called Principal Component Regression (PCR) is the sequential composition of a PCA step on the features followed by an OLS regression. In this case we use only the principal components with the highest variance to make predictions, by using the truncated scores as features in OLS. In some cases, however, a principal component with lower variance can have equal or even higher importance. This problem arises from the fact that the creation and selection of the principal components do not depend on the outcome value y ; they are selected an *unsupervised* way.

The most popular method applied in chemometrics is the Partial Least Squares (PLS) regression. In PLS the selection of the latent variables is a supervised procedure. The method projects the features and the prediction target or targets to new spaces with constrained dimensionality [66]:

$$\begin{aligned} X &= TP^{\top} + E, \\ Y &= UQ^{\top} + F, \end{aligned}$$

with the optimization criteria to maximize the covariance between these derived variables:

$$\max \text{cov}(T_i, U_i)$$

Having these representations, the method finds a regression model between these two spaces:

$$U = TD + H$$

Because T_i and U_i are corresponding latent variables with the highest covariance, this regression problem falls back to independent univariate problems: the D matrix we search for is diagonal.

As the OLS can be interpreted as maximization of the correlation, while the PCR selects latent variables according to the maximal variance criterion, PLS is a trade-off between these two cases [67].

The more general case of PLS briefly discussed above is called the PLS2, which can regress for several outcome variables together. This can help to improve predictions compared to building separate regression models. This principle is called multi-task learning in the machine learning literature [68]. We will use the same effect in matrix

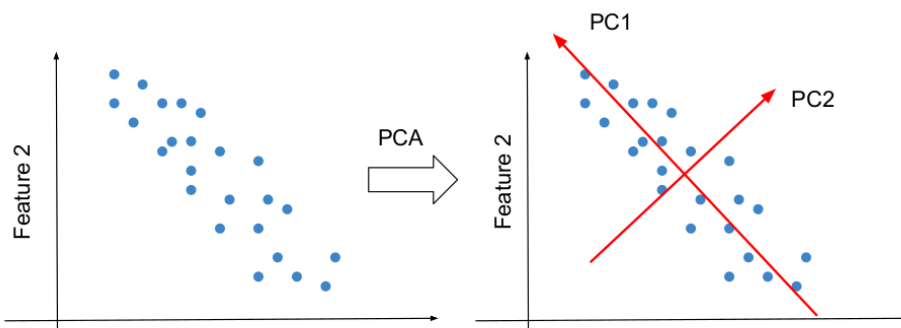


Figure 11 - Illustration of Principal Component Analysis (PCA) on the case of two strongly dependent features.

factorization models (see Section 3.17). If the prediction of only one outcome is needed, the PLS algorithm simplifies to a variant called PLS1 [69].

3.12 Basics of kernel methods

All of the techniques discussed above are linear, which means a given feature can have only an additive effect to the prediction. In some cases a better prediction can be made if we take into account their nonlinear effects and the interactions between different features. The easiest way to achieve this goal is to derive nonlinear combinations of the original covariates and use them in the regression procedure. A modern „off the shelf” method to derive a nonlinear counterpart for a linear method while preserving its favourable properties is *kernelization*. In this case we use the observation that the product $X^T X$ and the product XX^T contain the same information for modelling for a given sample by feature matrix X . While the size of the square matrix $X^T X$ is the number of features and the size of XX^T is the number of samples, the rank of the two matrices is the same. In the case where we have a very high dimensional feature set for moderate amount of samples using the later representation is more economical. This is the case when we use chemical fingerprints for approved drugs, or we derive a large number of nonlinear interactions of features. It is important to note, however, that this trick alone will not prevent the problems arising from the high dimensional feature set discussed above. We need to apply regularization to address this problem. The simplest illustrative case of the *kernel trick* is the kernelization of the ridge regression (a regularized form of Ordinary Least Squares) [70]:

$$(X^T X + \lambda I)\beta = X^T y.$$

As β is a vector with the size corresponding to the number of features, we can define it as a function of a new set of variables α with a size corresponding to the number of compounds as:

$$\beta = X^T \alpha.$$

Substituting back, and applying the Woodbury matrix identity for positive semidefinite matrices we get [71]:

$$\alpha = (XX^T + \lambda I)^{-1}y.$$

And to compute the prediction for a new input x , we need to calculate:

$$y_{\text{new}} = \alpha X^{\top} x_{\text{new}}.$$

On this point we can substitute X by a matrix Φ which is derived from the original features

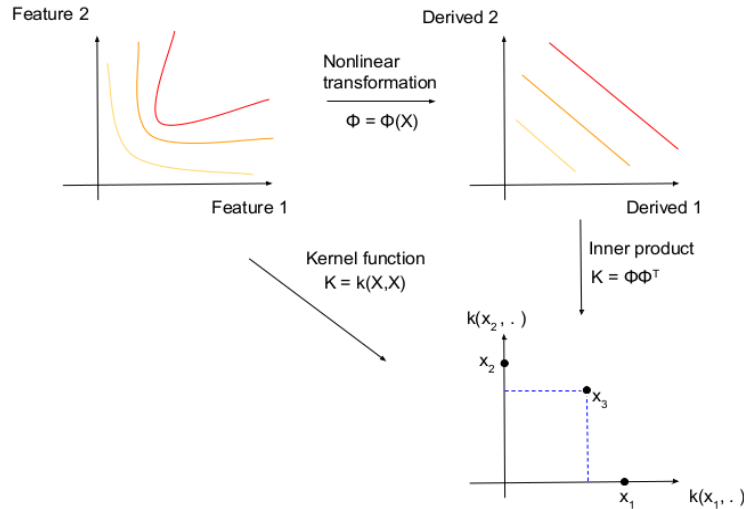


Figure 12 - Illustration of the kernelization. Using the kernel function $k(x_i, x_j)$ we can have a direct mapping from the input space to kernel space even if the dimensionality of our feature space is extremely high.

by nonlinear transformation $\Phi = \Phi(X)$. As there is no X outside of an inner product we can use $K = \Phi\Phi^{\top}$ everywhere. With the kernel trick we can even use an infinite number of features if we can directly compute the elements of K , which is always a finite sized sample-by-sample matrix (Figure 12). This matrix is called *kernel matrix*, and the direct mapping from the *input space* to the elements of K is called the *kernel function* $k(x_i, x_j)$.

A similarity relation discussed in the introduction, to form a valid kernel function, needs to meet some mild criteria. Because K is a symmetric matrix in the form $\Phi\Phi^{\top}$, the matrix calculated with the given similarity relation should be symmetric as well, and it should be factorizable in this form.

According to *Mercer's theorem* K is a valid kernel if it is symmetric and positive-semidefinite, that is

$$\forall g \in \mathbb{R}^n \quad g^{\top} K g \geq 0.$$

For an illustration, let us use the following kernel function:

$$k(x, y) = (x^\top y + 1)^2.$$

It can be shown using the original form of Mercer's theorem that this function is a valid kernel function (for the theorem and its application, see detailed description in [72]). To illustrate the kernel trick instead, let us expand kernel function above as

$$\begin{aligned} k(x, y) &= \left(\sum_{i=1}^F x_i y_i \right) \left(\sum_{j=1}^F x_j y_j \right) + 2 \sum_{i=1}^F x_i y_i + 1 \\ k(x, y) &= \sum_{i=1}^F \sum_{j=1}^F x_i x_j y_i y_j + 2 \sum_{i=1}^F x_i y_i + 1 \end{aligned}$$

from which we can read the feature map Φ :

$$\phi(x) = \langle x_i^2, \sqrt{2}x_i x_j, \sqrt{2}x_i, 1, \text{ where } i, j \in [1, F], i \neq j \rangle$$

The generalization of this kernel for a general exponent d , is called the *polynomial kernel*. Another, probably the most frequently used kernel in the machine learning literature is the radial basis function (RBF) kernel has the form

$$k(x, y) = e^{-\gamma \|x-y\|^2}.$$

It can be shown using Taylor-expansion that this kernel corresponds to an infinite dimensional feature space:

$$\phi(x) = \left\langle \frac{(2\gamma)^{\frac{k}{2}} e^{-\gamma \|x\|^2}}{\sqrt{k!}} x^k, k = 0, \dots, \infty \right\rangle.$$

As the above equation shows, all elementwise powers of x are included in the feature set, but the multiplier decreases strongly with increasing k .

3.13 Data fusion with kernel methods

To understand the complex nature of small molecule–target interaction, we need to use several type of data sources including chemical structure, mechanism, known targets, gene expression, known phenotypic effects; and integrate the information content of these sources. The different possible representation of these type of data – like different chemical fingerprints for the chemical structure - , and the diversity of the similarity

relations make the number of possible combinations even higher. Similarly to genomics, the data fusion in chemoinformatics and chemogenomics becoming more and more important [73].

It is practical to divide data integration methods into three categories, namely *early*, *intermediate* and *late integration* methods [73]. In case of early integration the feature vectors from different information sources are concatenated to a single vector, and the modelling procedure uses this as an input. If the model takes into account the correlation between the features, the model will use any interaction between features even between sources.

In case of intermediate fusion, the similarities of the entities are combined – usually added up - to form a fused similarity matrix. It is, therefore, called as kernel combination. In this case the between-source correlations are not taken into account. We generally assume that the within-source correlations are more important, therefore we may want to restrict the descriptive power of our model this way. The similarity of this motivation to the one behind dropout, a widely used technique in modern machine learning, would be worth further investigation [74].

In case of late fusion, separate models are trained based on the different data sources, and the decisions of the models are fused. Two possible fusion options are the score fusion, when the output of the models are directly combined, and the rank fusion. As we already discussed in the case of virtual screening, if we reduce our scores to ranks and then we fuse the ranked lists to a consensus ranking, we can combine outputs with significantly different score distributions.

Multiple Kernel Learning (MKL) is a commonly used technique for data integration, which depends on the application of the kernel trick. As an intermediate fusion technique, MKL depends on the linear combination of kernels:

$$K = \sum_k d_k K_k.$$

Here K_k is a kernel derived from the information source k , and d_k is a corresponding information source weighting which we want to determine during the model building. An intuitive way to think about the optimization of the data source weights is that we want

to create a combined similarity metric, which makes our query compounds as similar to each other as possible, while in the same time make the separation from the other class, or from the origin in case of one-class problem, possible.

We will follow the Lp-MKL formulation [75]. In case of a two-class classification problem the non-kernelized optimization objective, also called the primal objective, is:

$$\begin{aligned} \min_{w,b,\xi,d \geq 0} & \frac{1}{2} \sum_k w_k^\top w_k + C \sum_i \xi_i + \frac{\lambda}{2} \left(\sum_k d_k^p \right)^{\frac{2}{p}} \\ \text{s.t.} & y_i \left(\sum_k \sqrt{d_k} w_k^\top \phi(x_i) + b \right) \geq 1 - \xi_i. \end{aligned}$$

The objective contains three clearly identifiable contributions, namely the regularization of the model parameters, the classification error, and the regularization of the kernel

$$\max_{\alpha} \sum_i \alpha_i - \frac{1}{8\lambda} \left(\sum (\alpha^\top Y K_k Y \alpha)^q \right)^{\frac{2}{q}}.$$

weights. As the inequality constraint shows, the problem is equivalent to classifying in a concatenated feature space contrary to the early fusion, where we apply concatenation in the input space. Following the tedious derivation of *Sun et al.* we get the kernelized or dual objective as:

3.14 One-class Support Vector Machines

In its original form one-class SVM aims to identify a region of the input space where most of the training examples lie, or more precisely: to identify a function $f(x)$ which has a definite positive value $f(x) \geq 1$ in the region where a given $(1 - \nu)$ portion of the probability mass lies. In our application we will take advantage of the smoothness of $f(x)$ to prioritize the points, the candidate drugs, in the input space.

Let the training set be

$$x_i \in \mathbb{R}^M \quad i = 1, \dots, N.$$

Using the nonlinear transformation Φ we map the training samples to the feature space, where we search for the solution as the form:

$$f(x) = w^\top \Phi(x),$$

where the optimization goal is to make $f(x)$ positive for all training samples if it is possible, given some regularization constraint. The problem is equivalent to finding a separating hyperplane between the training samples and the origin in the feature space.

$$\begin{aligned} \min_{w, \xi, \rho} \quad & \frac{1}{2} \|w\|^2 + \frac{1}{\nu N} \sum_i \xi_i - \rho \\ \text{s.t.} \quad & w^\top \Phi(x_i) \geq \rho - \xi_i, \quad \xi_i \geq 0. \end{aligned}$$

It can be shown that we can write $f(x)$ in the following form:

$$f(x) = \sum_i \alpha_i k(x_i, x) - \rho,$$

where some of the α values will be zero. The support vectors – training examples with non-zero α values - will lie on the boundary of the set, and will give the ranking a max-score like behaviour, as pointed out in Section 3.3. It is also clear that ρ is only an additive constant, therefore irrelevant in the ranking case.

For a similarity metric S for which $S(x_i, x_j) \geq 0$ and $S(x_i, x_i) = 1$, the points lie in a hypersphere, and also in the same orthant. For illustration in the two dimensional case see Figure 13.

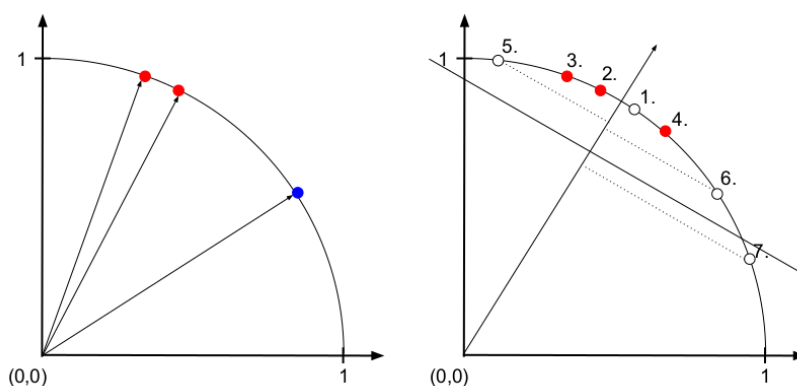


Figure 13 - Structure of the feature space in case of a normalized similarity metric (left), and the example of the prioritization using the separating hyperplane between the training compounds (red dots) and the origin (right).

A multiple kernel version of the one-class SVM can be derived analogously to the two-class classification case discussed above, from the primal optimization problem [10]:

$$\begin{aligned} \min_{w, \rho, \xi, d \geq 0} \quad & \frac{1}{2} \sum_k w_k^\top w_k - \rho + \frac{1}{\nu L} \sum_{i=0}^L \xi_i + \frac{\lambda}{2} \left(\sum_k d_k^p \right)^{\frac{2}{p}} \\ \text{s.t.} \quad & \sum_k \sqrt{d_k} w_k^\top \phi(x_i) \geq \rho - \xi_i \end{aligned}$$

Leading to a prioritization score:

$$f(x) = \frac{\sum_i \alpha_i \sum_k d_k K_k(x_i, x)}{\sqrt{\sum_k d_k \alpha^\top K_k \alpha}}$$

3.15 Semi-supervised and Positive and Unlabelled Learning

The SVM models provide good predictive performance [76], but there is space for improvement. Like in the case of similarity searching, performance can be improved by introducing the information of the input distribution, even if this information lacks labels or known outcome values. In machine learning, when a method solves a supervised learning task utilizing also the unlabelled examples available is called semi-supervised learning. An extreme case of semi-supervised learning when only positive labels are available is called positive and unlabelled (PU) learning [77]. This means that in our case only active compounds are labelled for an indication, every unlabelled compound can be an undiscovered active one, or an inactive.

The two most well-known semi-supervised learning metaheuristics are self-training and co-training. In case of self-training a classifier is trained on the labelled dataset, and then prediction is computed for all unlabelled samples. In the further iterations highly confident predictions from the previous iteration are added to the real labelled set, and the classifier is re-trained [33]. An application of this metaheuristic in chemoinformatics is turbo similarity searching [32]. The main drawback of this method is that a false prediction can be self-reinforcing. Co-training is a similar method using two different classifiers training each other with highly confident predictions.

We can view the iterative application of a prioritization method by a human expert as a kind of semi-supervised learning assisted with weak evidences or expert knowledge [14]. In this scenario a prioritization framework is used to rank chemical compounds, and then a human expert selects some of the high ranked candidates into a query set in the next iteration.

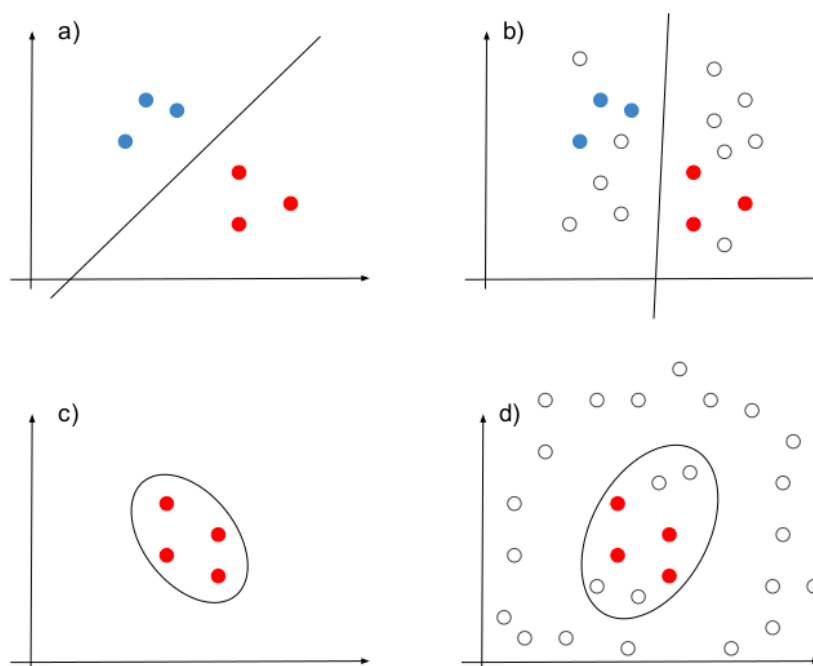


Figure 14 - The effect of the utilization of information about the unlabelled samples to the decision boundaries of the models. Two class supervised learning task (a) versus semi-supervised classification (b), and one-class learning (c) versus PU learning (d).

Placing the decision boundary the way that it corresponding to our structural assumption of the input distribution can lead to an increase in model performance. For example this illustration relies on the assumption that our class is compact.

In case of PU learning if the assumption holds that the missing labels are probably unobserved negative labels, we can label a random set as negative. We should be careful, however, if we would like to predict new positive cases.

An illustration of the semi-supervised learning concept is shown on Figure 14. Based on our *a priori* assumption about the distribution conditioned on the label, we can use the empirical input data distribution to choose a better classification model.

3.16 Generalization error and cross validation

Generalization ability is an important property of a prediction method because we want to use our models in new chemical libraries. If a dictionary of known input–output pairs is used as a classifier, it is obvious that the performance on this original training dataset will be perfect, given that our training data is perfect. For example, a dictionary-based QSAR method would just search for the given chemical structure in the dictionary, and output its measured activity. It is easy to see that this type of method is totally useless to predict new chemical series. This type of non-flexible behaviour is called *overfitting*.

Cross-validation is a model validation technique for assessing future performance on independent datasets [78]. To avoid overfitting, we can split our dataset to a training set, and a non-overlapping test set. The latter is not used in model building, and it is only used to compare the prediction of the model with the measurement values. One of the most common methods is *n-fold* cross validation. In that case we divide our dataset to *n* equal sized folds, and use one of them as test set, and the rest as training set. We build the model with the *n* possible training sets separately, and evaluate it with the corresponding test set. Using the evaluation metrics we can compute statistics like mean performance, or assess the statistical significance of the performance differences. The illustration of *n-fold* cross validation is shown on Figure 15.

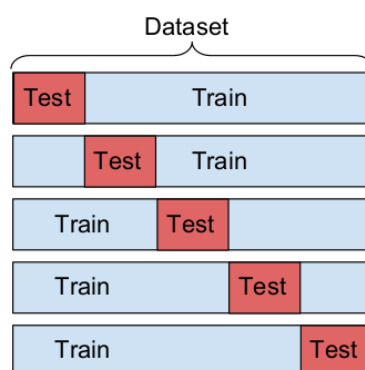


Figure 15 - Illustration of 5-fold cross-validation. The dataset is partitioned to five parts four out of which is always used as training set and one as a test set.

A simpler to implement version is bootstrap cross-validation when we just sample a portion of the samples to the test set, and train the model on the rest of the data, and repeat

it several times. In this case the variability of the result can be higher because the test sets may overlap, so more repeats need to be calculated to reach the same statistical power.

If the model building has hyperparameters, like ν or C in case of SVM, or the number of layers in case of a neural networks, we need to tune these hyperparameters. If we want to assess the predictive accuracy on an unrelated dataset we cannot do it by just optimizing the test set performance. If we did it, the hyperparameters would be tuned specifically to reach good performance on the test set and then we would overestimate the generalization performance. Therefore we need to use nested cross-validation. If we are only interested to compare different parametrizations or different methods, but we do not want to claim a quantitative measure of accuracy in a practical situation, we do not need external folds.

In nested cross-validation (see Figure 16) we partition the dataset to n outer folds, one out

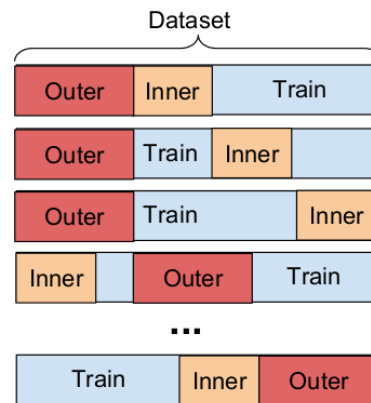


Figure 16 - Illustration of 3x3-fold nested cross-validation. The dataset is partitioned to three equal parts two out of which form the outer training set in every iteration. This set is now partitioned again to three equal parts, and analogously an inner training set, and a validation set is formed.

of which is used as test set for measuring the performance. The training set is then partitioned to m inner folds, one out of which is used for test set during the hyperparameter tuning often called validation set, and the others as training set.

3.17 Macau: Bayesian Multi-relational Factorization

As it is discussed above, if we do not have enough samples to properly identify a model, we should determine the *a posteriori* distribution of the models instead (Section 3.7). In the Bayesian framework we can control model complexity through the prior by simply

making $P(M)$ small for complex models and large for simple ones. However, we can still use other regularization techniques, like dimensionality reduction. Macau is an example for this as a Bayesian matrix factorization method for large scale incomplete matrices with high dimensional side information. Macau was developed in the STADIUS bioinformatics research group at KU Leuven with my participation. In that project I was responsible for model specification and for evaluation in pharmaceutical applications. It is a general tool that is designed with the special requirements of compound-protein interaction prediction task in mind. The so called side information is composed of additional features for rows and columns. If the rows of the matrix correspond to chemical compounds, side information can be the set of chemical fingerprints. The likelihood of the observations $P(D/M)$ have the following form in this specific case:

$$P(Y|U, V, \alpha) = \prod_{(i,j) \in I} \mathcal{N}(Y_{ij} | u_i^\top v_j, \alpha^{-1}),$$

where U and V variables have their own prior probability, corresponding to the general term $P(M)$. The rows and columns of Y can be regarded as entities like drugs and indications, and the matrix encodes relations between them. These entities have their own representations as a form of the vectors u_i and v_j . Every entity, like every compound or indication category, has its own descriptor vector with the length of K , where K is a small number relative to the size of the matrix. Therefore, we limit the number of free variables from $N \times M$ to $N \times K + M \times K$, where $K \ll N, M$.

Macau has a hierarchical structure, and can be described with a graphical model, which is similar to the Bayesian networks discussed before (see Figure 17) [79, 80]. In the following discussion the precision of the observations (α) is assumed to be known *a priori*,

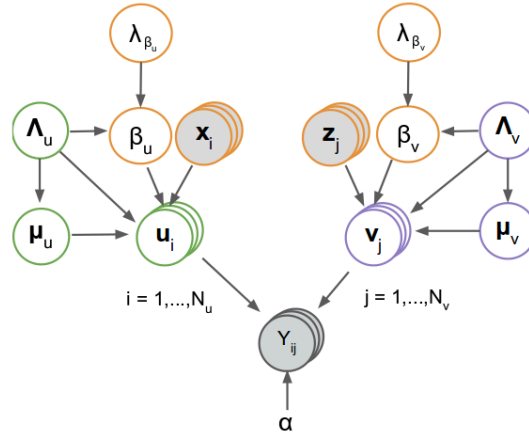


Figure 17 – Probabilistic graphical model of Macau. The graph shows the hierarchical structure of the model: the prior of u_i and v_j depend on the row and column side information x_i and z_i respectively, and link matrices β_u and β_v are learned for the row and column side information. All unobserved -white- variables have their own priors.

but in a more general form of the model it can have a Gamma distribution as non-informative prior.

The model can be defined by giving the local probability models for every variable similarly as $P(Y|U,V,\alpha)$ was given above. As the model is perfectly symmetric, we will give the formula only for one of the variables. In the second level the prior of the latent variables is defined as:

$$P(u_i | \Lambda_u, \mu_u, x_i, \beta_u) = \mathcal{N}(u_i | \mu_u + \beta_u x_i, \Lambda_u^{-1}).$$

As x_i and x_j are observed variables, in the third level we are left with two vectors of mean (μ_u, μ_v) , two precision matrices $(\Lambda_u$ and $\Lambda_v)$ and two link matrices $(\beta_u$ and $\beta_v)$. The speciality of Macau is the proposed scale invariant prior over the link matrices:

$$P(\beta_u | \Lambda_u, \lambda_{\beta_u}) = \mathcal{N}(\text{vec}(\beta_u) | \mathbf{0}, (\Lambda_u \bullet \lambda_{\beta_u} I_{F_u})^{-1}),$$

where \bullet denotes the Kronecker product operation [81]. This prior is invariant to the scale of the latent variables. The determination of the probability distribution of the link matrix β_u can be regarded as solving a multivariate regression problem (see Section 3.11) in a Bayesian context, when our feature matrix is X , and our outcome matrix is V . As V is a latent variable and not an observed variable, our regression is “shooting a moving target”, therefore the introduction of an adaptive self-adjusting prior was necessary. The prior

over A_u and μ_u is a standard normal-Wishart distribution. Only one pair of variable is left for the fourth level: λ_{β_u} and λ_{β_v} . Their *a priori* distribution is the Gamma distribution.

We can use the independence relations encoded in the graphical model to determine the distributions of the variables. For example, if we knew the value of all variables in the Markov blanket of u_i (x_i, β_u, μ_u, A_u and Y_{ij} and v_j) we could compute the distribution of u_i . As we do not know the values of these variables, we use a Monte Carlo method called Gibbs-sampling to draw samples from the joint probability distribution of the variables [82].

4 Objectives

The objectives of my doctoral thesis are:

- To develop a novel data fusion method for the prediction of the biological effects of small-molecular drugs by integrating heterogeneous information sources.
- To apply the data fusion method for finding Parkinson's disease related drugs, and to evaluate the ability of this method to enhance drug discovery, especially drug repositioning.
- To develop and evaluate a novel matrix factorization based method capable of predicting multiple activities simultaneously, and to compare it with a single target baseline method.
- To adapt and apply a novel Bayesian multivariate statistical technique to identify genetic variants predictive of the interpersonal variability of methotrexate pharmacokinetics at high dose levels.

5 Methods

5.1 Information sources

At the start of the research information sources describing compounds were constructed: Molecular Access Keys (MACCS); molecular connectivity, shape and electrotopological fingerprint (MOLCONN-Z); 3D pharmacophore based fingerprint; side effect occurrences and frequencies; and known drug-target interactions. We define the vector representation of the compounds for each information source. Also similarity metrics was identified to compute pairwise similarity kernels from the features for the methods requiring similarities. The Tanimoto similarity was used for every information source with binary features, whereas the cosine similarity was applied for sources based on real valued features.

The basic summary provided below describes the source of the data, the software version used to generate the features and the number of drugs for which the given type of information is available. It also shows the mean and median value of all pairwise similarities and the histogram of all pairwise similarities, which gives an image of the distribution of similarity relations in the space defined by the given features.

Two main versions of these information sources were used during the work: the first version relies on the Anatomical Therapeutic Chemical Classification System (ATC) codes as identifiers for the compounds [10]. Because of the multiple occurrences of some compounds in the ATC hierarchy, in later publications we used a new version, where the identifiers are standardized English International Nonproprietary Names (INNs) of the compounds. The properties of these two datasets and the results based on them are qualitatively the same.

It seems to be a rational choice to use the chemical structure of the compounds as an identifier, but the possible salt forms and different tautomers make the mapping labour intensive, therefore in the case of approved drugs an identifier like INN is a more convenient choice.

Table 5 - Information sources used in the different phases of the work

	Target	Freq	Preval	3D	MACCS	Molconn.	TFIDF	Used ID
Method study (CMC)	X	X	X	X	X	X	X	ATC
Amantadine study (FMC)	X	X		X	X	X	X	INN
Parkinson's study (CTMC)		X		X	X	X	X	INN
Multi-target	X	X		X	X	X	X	INN

The target information source is special in a sense that it can bias the prioritization towards known targets. If we would like to be conservative, we can drop this information source to find out if our method can identify a target which is already known from the other sources (see Table 5). In the studies, where we compared two statistical methods, this bias is irrelevant because the extra knowledge can help both methods equally. An old version of side effect prevalence based data source (*Preval*) contained information only for approximately 100 drugs; we therefore decided to drop it from the second version of the dataset.

The pairwise overlap of the data sources is presented in Table 6. For every pair of data sources the number of drugs present on both data source is given. In the diagonal the size of the data sources are presented.

Table 6 - Overlap of the data sources: The table contains the number of drugs occurring in two data sources simultaneously. The diagonal elements are the sizes of the data sources.

	MACCS	MOLCONN	3D	FREQ	TARGET	TFIDF
MACCS	1851					
MOLCONN	1823	1823				
3D	1754	1753	1755			
FREQ	532	519	511	543		
TARGET	1087	1074	1055	404	1162	
TFIDF	868	853	819	513	766	925

MACCS: Molecular Access Keys (Schrodinger Suit 2012 Canvas)

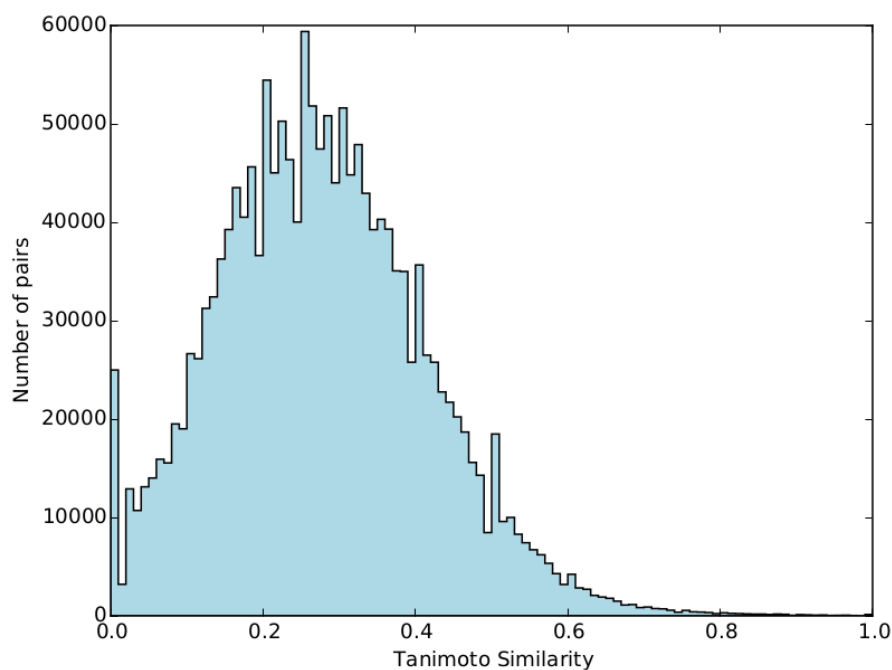


Figure 18 - Histogram of Tanimoto similarities based on MACCS keys (Number of drugs: 1851, Mean similarity: 0.2786, Median similarity: 0.2708)

It is a MACCS key based binary fingerprint, where all binary features directly correspond to a question about the existence of a structural pattern defined by a Smiles Arbitrary Target Specification (SMARTS) query and no hashing or folding is applied. In this work we used the standard MDL definition with 166 queries. The histogram of the pairwise Tanimoto similarities is presented on Figure 18.

MOLCONN-Z: Molecular Connectivity, Shape and Electrotopological fingerprint
(Schrodinger Suit 2012 Canvas)

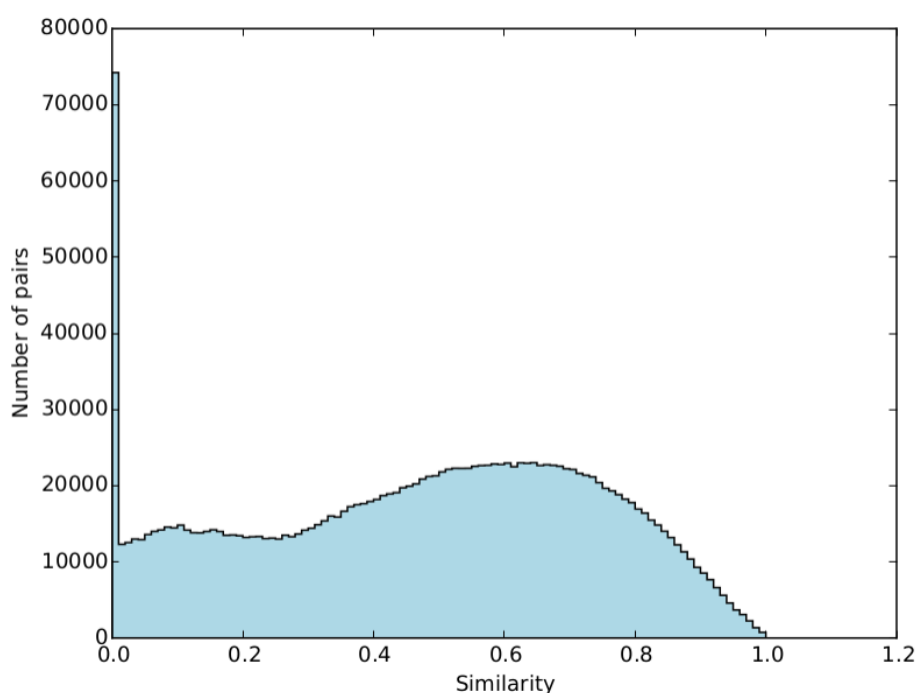


Figure 19 – Histogram of cosine similarities based on the MOLCONN-Z descriptor.

(Number of drugs: 1823, Mean similarity: 0.4720, Median similarity: 0.5000)

We calculated the Molconn-Z electrotopological state (Estate) with all four options (Key, Count, Sum, Average) available in Schrodinger Canvas software, and concatenated the result to get a feature vector with maximal length of 352 for all compounds. The histogram of the pairwise cosine similarities is presented on Figure 19.

3D pharmacophore based fingerprint (Schrodinger Suit 2012 Canvas)

The fingerprint is generated from triplets of pharmacophoric features and their distances. The conformers used for the analysis were generated during the fingerprint calculation process with default parameterization. The histogram of the pairwise Tanimoto similarities is presented on Figure 20.

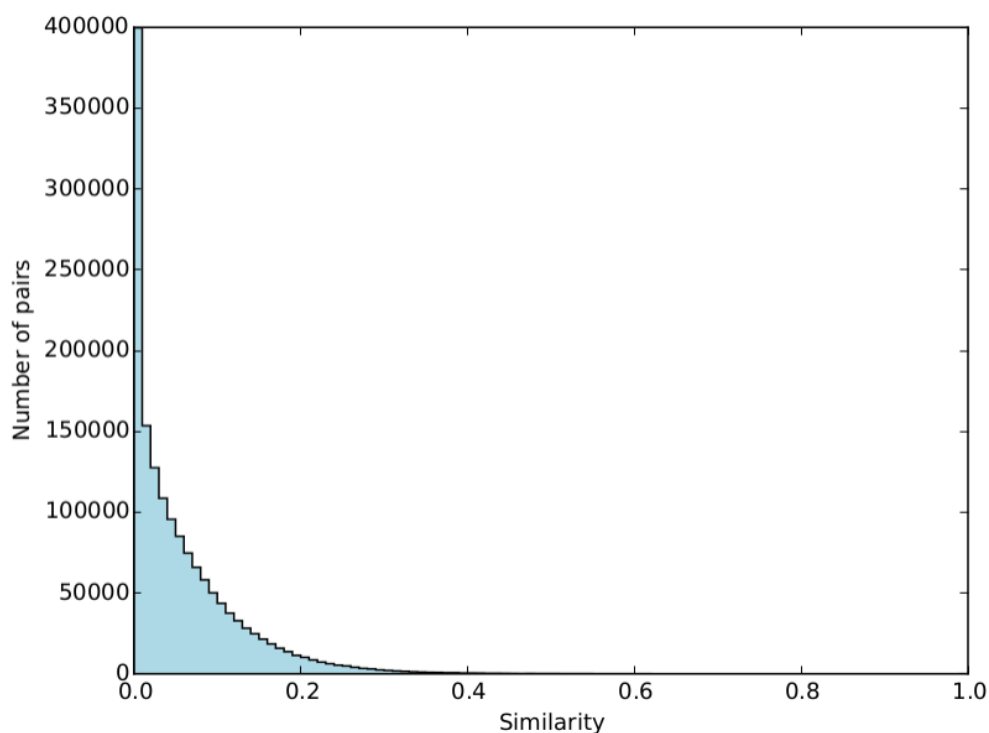


Figure 20 – Histogram of Tanimoto similarities based on three dimensional pharmacophore fingerprint (Number of drugs: 1755, Mean similarity: 0.0380, Median similarity: 0.0600)

FREQ: Side Effect Frequencies

This fingerprint was built based on the data we extracted from the SIDER database [51]. Every real valued feature corresponds to a side effect, and the value between 0 and 1 measures the prevalence of this side effect in the treated population. The histogram of the pairwise cosine similarities is presented on Figure 21.

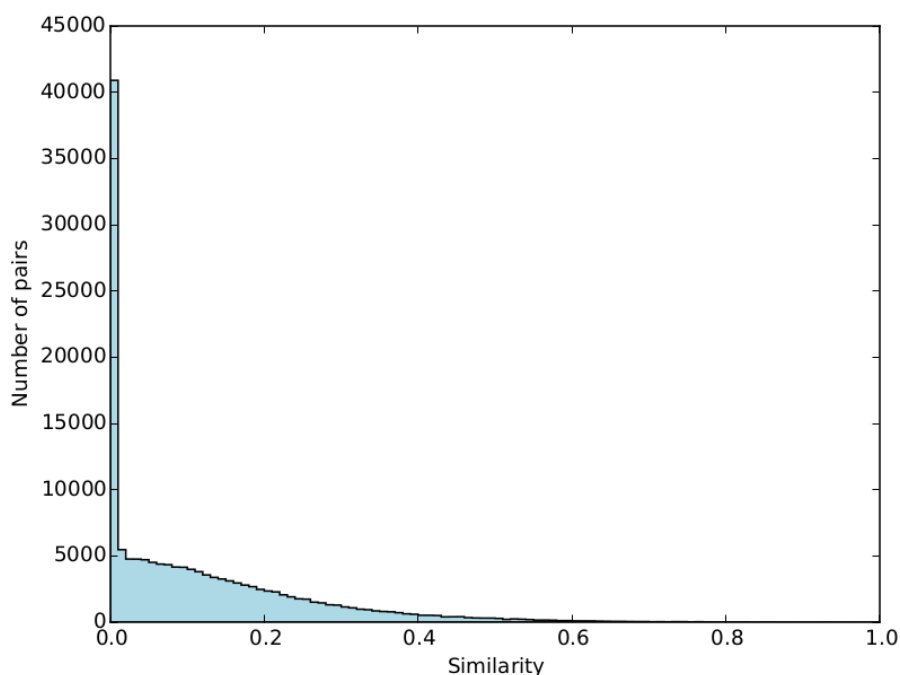


Figure 21 - Histogram of cosine similarities based on side effect frequencies (Number of drugs: 543, Mean similarity: 0.1195, Median similarity: 0.0794)

TARGET Known Drug-target interactions

A binary descriptor based on validated targets of the drug, extracted from the DrugBank database [83]. Every feature corresponds to a biological target. Because the number of validated targets for a given drug is usually very small, even if the compound in practice can be quite promiscuous, these vectors are very sparse.

Table 7 - Statistical properties of the pairwise Tanimoto similarities based on the Target data source

Number of drugs: 1162	Tanimoto similarity	
	Zeros removed	Zeros not removed
Mean similarity	0.3146	0.0082
Median similarity	0.2000	0.0000

Because of the sparseness of this relation, histogram is dominated by a peak at 0.0 similarity level. Mean and median similarity calculated based only on the nonzero values (see Table 7).

TFIDF Side effect related terms

This one is a continuous valued descriptor, where each position corresponds to a relevant term and its value is the *tf-idf* score of the term in the package leaflet corpus. We used documents from the DailyMed database, which contains package leaflets submitted to the FDA [84]. These labels are stored in a standardized semi-structured XML format. They contain information about the active substances, manufacturer, indications, dosage, contraindications, possible drug interactions and side effects among others.

To compute *tf-idf* score, first we need to compute the term frequency:

$$tf_{ij} = \frac{n_{ij}}{d_j},$$

where n_{ij} is the number of times term i appears in the document j , and d_j is the length of document j in words. Here document j corresponds to the package leaflet of drug j . As a next step we need to compute the inverse document frequency, which measures how informative, in other words how specific, a term is in general:

$$idf_i = \log \left(\frac{N}{n_i} \right),$$

where n_i is the number of the documents containing the term i , and N is the number of all documents. It is clear that if all documents contain a word, that word has very little information about the drugs. The *tf-idf* score is the product of tf_{ij} and idf_i .

We used the MedDRA (Medical Dictionary for Regulatory Activities) to create a dictionary of side effects in the form they are used in package inserts [85]. MedDRA is a standardized, international, officially adopted terminology to facilitate the sharing of regulatory information. It has a tree structure with five specified levels: System Organ Class (SOC), High Level Group Term (HLGT), High Level Term (HLT), Preferred Term (PT), and Lowest Level Term (LLT). Only PTs and LLTs were used to create this information source. Every position in a descriptor corresponds to a PT, and every LLT occurrence in the corpus was counted to the corresponding PT. For example the LLT *Joint inflammation* corresponds to the PT *Arthritis*.

We filtered these terms further using the UMLS (Unified Medical Language System) ontology [86], using only terms that are assigned for one of the following four UMLS semantic types:

- Anatomical Abnormality
- Finding
- Natural Phenomenon or Process
- Sign or Symptoms

Because MedDRA is also part of the UMLS system, the filtering is directly applicable.

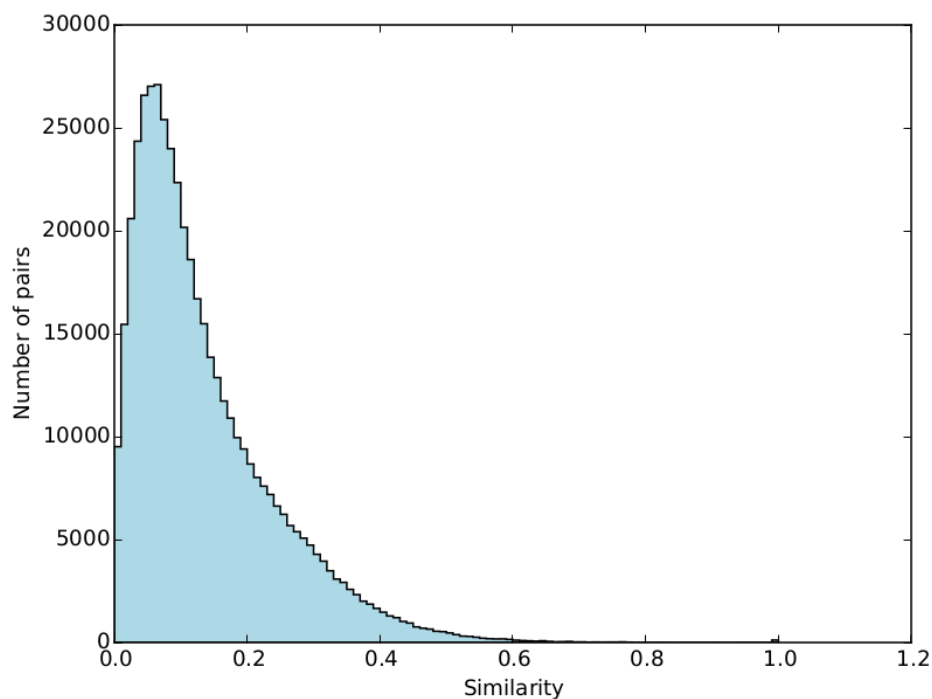


Figure 22 - Histogram of cosine similarities based on side effect $tf-idf$ scores in package leaflets. (Number of drugs: 925, Mean similarity: 0.1364, Median similarity: 0.1057)

Finally, a descriptor vector is formed for each drug from all $tf-idf_{ij}$ scores corresponding to that drug id j . The histogram and the statistical properties of the pairwise cosine similarities are presented on Figure 22.

5.2 Redundancy and complementarity of the information sources

To assess the common information content of these data sources, that is to evaluate their complementarity, we computed the Spearman correlations of all pairwise similarities (Table 8). Because the distribution of the similarities is very different kernel to kernel, the correlation of the ordering of these similarities is more suitable than a standard Pearson correlation. As it is discussed earlier, this ordering is equivalent to a quantile normalization approach, which maps the different empirical distributions to a uniform distribution.

Table 8 - Spearman correlations between pairwise similarities based on different data sources using the ATC based kernels (: $p < 10^{-5}$; **: $p < 10^{-10}$; ***: $p < 10^{-20}$)*

$\alpha = 0.001$	Target	Sider	Prev	3D	MACCS	TF-IDF	Molconn
Target	-	0.4763	0.4927	0.4911	0.4941	0.4837	0.4743
Freq	***	-	0.1996	0.067	0.0637	0.1465	0.0006
Prev	***	***	-	0.1377	0.0762	0.2543	-0.0412
3D	***	***	*	-	0.3798	0.0934	0.2764
MACCS	***	***	**	***	-	0.1343	0.4433
TF-IDF	***	***	***	***	***	-	0.0743
Molconn	***	not sig.	not sig.	***	***	***	-

5.3 Evaluation framework for the fusion methods

I participated in the development of the novel Kernel Fusion Repositioning (KFR), which method uses the one-class SVM framework and serves as a reference model class in the comparison of other data fusion methods. As the one-class SVM seems to be quite

insensitive to the parametrization in a prioritization setting, we just used a fixed parameter $v=0.4$ for all of our experiments. In the late fusion setup we computed prioritization based on different data sources separately, and fused the ranking with the Borda protocol. We used the Lp-MKL formulation for the intermediate fusion, and we used an in house implementation of the SMO-MKL solver by *Sun et al.* extended with the one-class option [75].

AUC[ROC], AUC[CROC(exp)], BEDROC and fixed threshold *sensitivity* and *specificity* measures were used to evaluate predictive performance. The early discovery focus was $\alpha=20.0$. Two thresholds were introduced for both the sensitivity and the specificity: top25 and top100. We predicted the membership of Level 4 ATC classes and evaluated the performance with bootstrap cross-validation: 30% of the class members were randomly selected as a test set, and 70% were kept for training, using 100 repetitions.

5.4 Drug-Indication reference set

To evaluate the different data fusion methods we need a drug classification system which is widely accepted and defines the “gold standard” indications. ATC, a widely accepted classification system was utilized to compare the predictive performance of the different ranking methods. ATC is a five-level taxonomy maintained by the Collaboration Center for Drug Statistics Methodology of the World Health Organization. The first level, called the anatomical main group, is the most general group, based on the organ or system on which the drug acts, like dermatologicals or nervous system. The second level, called the therapeutic main group, indicates therapeutic categories like antihypertensives, immunosuppressants or analgesics. The third level can indicate therapeutic or pharmacological subgroups, like antidepressants, or opioids. The fourth level can indicate chemical, therapeutic or pharmacological subgroups, like sulphonamides, or selective serotonin reuptake inhibitors. It is important to note that the same compound can appear multiple times in the taxonomy if it has more than one indication. For example the macrolide tacrolimus has two different ATC identifiers: D11AH01 and L04AD02. The former corresponds to D (dermatologicals), 11 (other dermatological preparations), AH (non-corticosteroid agents for atopic dermatitis), while the latter is L (immunomodulators and antineoplastics), 04 (immunosuppressants), AD (calcineurin inhibitors).

In the experiments we used the 95 Level 4 ATC classes from our dataset, which contained at least 6 drugs, without eliminating duplicated ATC identifiers. We omitted 6 categories because of their inhomogeneity: *Other ophthalmologicals* (S01XA), *Detoxifying agents for antineoplastic treatment* (V03AF), *Antidotes* (V03AB), *Other nasal preparations* (R01AX), *Other plain vitamin preparations* (A11HA), *Other antineoplastic agents* (L01XX), *Other dermatologicals* (D11AX), *Electrolyte solutions* (B05XA), and *Other cardiac preparations* (C01EB).

5.5 Application for Parkinson's disease therapy

Parkinson's disease (PD) is one of the most well-studied neurodegenerative diseases characterized by the progressive loss of dopamine producing neurons in *substantia nigra*. As the research group at the Department of Organic Chemistry has interest in Parkinson's disease therapies, we applied the data fusion based methodology to prioritize repositioning candidates for Parkinson's disease (PD). [14]. Nevertheless, the developed methodology can be applied to a wide range of repositioning projects in general. According to our current knowledge all neurodegenerative diseases share, with different levels of importance, the following underlying mechanisms: oxidative stress, neuroinflammation, mitochondrial dysfunction, protein misfolding and aggregation, glutamate excitotoxicity, proteosomal dysfunction, disrupted intracellular transport and neurofilamental network, microglial activation and abnormal apoptotic behaviour [14, 16].

To apply the methodology on practical pharmacology problems, there are some common steps to be done [14]. These steps are the following:

1. Definition of the broader prioritization goal

The prioritization can be a single run or it can be a sequential process. The goal can be to find drugs for an indication, or to find an indication to a drug. In our specific application, the goal was to find FDA approved drugs with good repositioning potential as a PD therapy.

2. Construction of the candidate list

We used the entire set of approved drugs as a candidate set. We could use any set, like a proprietary chemical library, or a subset of approved drugs. For example, as we search candidates for a central nervous system (CNS) related indication, we can pre-filter the candidates based on their blood-brain barrier penetration ability. Other options are filtering based on intellectual property considerations, toxicity related substructures or unwanted biological effects.

3. Construction of special kernels

As discussed earlier, the similarity of compounds can be assessed outside of the classic chemical representation space. One approach is the side effect based similarity, first applied by *Campilos et al.* and further discussed in one of our publications [10, 52]. Other rich sources of information can be constructed from chemically perturbed gene expression profiles, like from the CMAP or LINCS datasets [42, 87]. A use case for the application of CMAP profiles is discussed in detail in our work [12]. Another promising option is the incorporation of disease specific information sources, and expert knowledge through kernels. An interesting new possibility is the data source construction from High Content Imaging (HCI) screens [88].

4. Design and construction of the query

An important property of the query is heterogeneity, as it is also discussed in the case of group fusion [26]. To a given limit, heterogeneity is desirable as it increases the probability of non-trivial hits. Too high heterogeneity on the other hand can lead to anomalous behaviour. We constructed four different queries representing four subcategories or mechanisms of action such as neuroprotective agents, dopaminergic agents, muscarinic agents, and NMDA antagonists (see Table 9). Designing a query for prioritization is equivalent to setting the focus of the *in silico* study. Intuitively, we need to describe the indication we are interested in with a set of compounds. According to our studies, an optimal query size is around 3-10 compounds, but the query size can deviate significantly from this value in special cases.

Table 9 – The four Parkinson’s disease related queries with their descriptions.

Query	Description
amantadine pramipexole rasagiline	Neuroprotective agents: Agents with disease modifying effect and the ability of slowing or reversing disease progression.
bromocriptine cabergoline pramipexole rotigotine	Dopaminergic agents: Direct agonists of various dopamine receptors replacing the effect of the missing endogenous ligand.
amantadine budipine ifenprodil memantine	NMDA antagonists: Antagonists of the N-methyl-D-aspartate sensitive ionotropic glutamate receptor, believed to protect against glutamatergic excitotoxicity [89].
benzatropine biperiden trihexyphenidyl	Muscarinic antagonists: Agents used for reducing the relative cholinergic hyperactivity in the central nervous system caused by dopamine deficiency, restoring the striatal dopaminergic-cholinergic balance.

5. Running the method and evaluating the performance

There are diagnostic steps which can be done to rule out meaningless results. Checking the query heterogeneity – e.g. ISS/UAS value - before the run is the first diagnostic step. Checking the positions of the query compounds in the output list is also informative. If some candidate compounds got higher rank than some query compounds it can signal a strong hit, but if too many candidates had been ranked before the query it is a strong signal of extreme heterogeneity.

6. Extracting knowledge from the ordering

There are several ways to extract information from the resulted ordering, in addition to the investigation of the top hits. One option is to apply filters to lower the number of compounds we need to investigate. These filters can be chemical structure based or text mining based filters. We applied a PubMed based filter, where we filtered out compounds without co-occurrences with the terms “*PD*”, “*Parkinson*” or “*Parkinson’s disease*” in PubMed abstracts. Other options are filtering based on physicochemical properties, or based on functional group occurrences [90].

In addition to the filtering we can use enrichment analysis to test, if there is a property enriched in the top of the list. The application of the compound set enrichment analysis (CSEA) is discussed in our publication [12]. Compound set here means compounds having common properties interesting for our purpose, like common mode of action, target, indication or side effect [12]. The idea originates from gene set enrichment analysis (GSEA) [91].

We used the SaddleSum algorithm for enrichment analysis [92]. The intuition behind the algorithm is rather simple. We are interested in the enrichment of certain annotations on the top of our prioritization list. We have a vocabulary V of compound sets; the members of each set shares the same annotation. In our examples we will use the ATC Level 4 classes as annotations. We can collect a weight for every annotation by adding up the inverse rank or the score of all compounds sharing the given annotation.

To answer the question 'Is the given annotation significantly enriched on the top of the list?' we have to ask how likely it is that if we randomly pick entities, the sum of weights exceeds S . This probability will be our p-value. If this probability is low, it is highly likely that the enrichment is not caused by chance.

5.6 Evaluation of Macau

I also participated in the development of the Bayesian matrix factorization method Macau for the drug-indication prioritization task. An important aim of the present work is to make the method applicable for settings without negative samples. The probability that a missing association does not hold is much higher than the probability that it exists but it has not been verified yet. In the research a well-established strategy was selected to randomly choose a subset of the missing associations identified as the negative set.

Using the 4th level of the ATC hierarchy, we created a 718 x 99 matrix where the rows represent compounds, and the columns represent ATC Level 4 classes. The created matrix is sparsely filled with 872 ones with average of 1.21 classes per compound, which corresponds to a fill rate of 1.2%. We used the same ATC class level as in the one-class SVM experiments to make the interpretation of the results easier. There are, however, factors making the strict comparison difficult. First of all, the one-class SVM experiments were carried out using ATC codes as identifiers, while the new version of the kernels uses INNs. Secondly, we use Macau as a least-square classifier in the PU learning setting discussed earlier, while the one-class SVM does not use any information about the non-labelled compounds. Finally, Macau is capable of predicting multiple targets simultaneously.

We left out the compounds for which the side information is unavailable in every experiment, therefore the real factorized matrix is somewhat smaller than 718 x 99 (See Table 10).

Table 10 - The sizes of the matrices in the different Macau runs.

Data source	Drug-ATC class matrix		Drug-Feature Matrix	
	Number of drug	Number of ones	Num. of features	Nonzero features
MACCS	623	774	152	24166
MOLCONN	617	768	348	22332
3D	597	748	49710	387128
TFIDF	485	620	2339	59459
TARGET	534	658	1024	2080

As we mentioned above, to predict the missing elements in the matrix we need to include zeros with a prior probability. We chose to randomly add 4 times as many zeros as known

membership relations and repeated this imputation 20 times. To validate that the multi-task effect between ATC level 4 classes can improve our results we used a column-wise ridge regression (a form of regularized OLS regression) as a benchmark.

5.7 Analysis of the methotrexate pharmacokinetics

The second major topic of my thesis is pertaining to personalized medicine. Personalized medicine can help in the clinic by suggesting tailored therapies, and also in the pharmaceutical research, as it can facilitate more effective drug development. I participated in a research about analysing interpersonal variability of methotrexate pharmacokinetics at high dose levels in children with osteosarcoma. The aim of our study was to investigate possible genetic factors and their role in the inter-individual differences of the pharmacokinetics and toxicity of methotrexate.

Osteosarcoma is a primary malignant bone tumour with the highest prevalence in the group of children and young adults. One of the established therapies is high dose methotrexate chemotherapy. The reduced elimination of this drug can lead to toxicities, especially hepato- and myelotoxicity.

In the following I will describe a dataset collected at the 2nd Department of Paediatrics of Semmelweis University, and which I will use to demonstrate the application of the Bayesian multilevel relevance analysis in pharmacokinetics studies. As a member of the research group I analysed the effect of 29 preselected single nucleotide polymorphisms (SNP) from the genes ABCB1, ABCC1, ABCC2, ABCC3, ABCC10, ABCG2, GGH, SLC19A1, NR1H2 (see the details in Table 11). In gene selection we significantly relied on the literature and on relevant scientific findings as well. When estimating functionality we relied on the classification of the polymorphism and its localization. The SNPs were ranked from the highest to the lowest functionality as non-synonymous, localization in the promoter region, localization in the 3' UTR region, synonymous and intronic localization. Only polymorphisms with minor allele frequency (MAF) greater than 10% were selected taking care to cover the most haplotype blocks possible.

The isolation of the genetic material from blood was carried out by using Qiagen isolation kits (QIAmp DNA Blood Maxi Kit / QIAmp DNA Blood Midi Kit; Qiagen, Hilden,

Germany). For sequencing GenomeLab SNPstream genotyping platform (Beckman Coulter) was used.

Table 11 - Selected SNPs for genotyping (table adapted from [60]). We will concentrate on the methodology in this work, therefore only properties relevant for the statistical analysis are presented. Detailed description of the biology can be found in our publication or in the doctoral thesis of Dr. Marta Hegyi [60, 93].

Gene	SNP	Alleles	N11 (%)	N12 (%)	N22 (%)	MAF (%)	HWE
ABCB1	rs1045642	C/T	16	23	15	49	0.27
	rs1128503	C/T	16	29	14	48	0.28
	rs9282564	A/G	31	10	0	12	0.37
ABCC1	rs4148358	G/A	32	11	2	17	0.42
	rs246219	G/A	38	9	1	11	0.59
	rs246221	A/G	25	17	5	29	0.42
	rs12922588	A/G	20	23	9	39	0.49
	rs215060	A/G	28	18	0	20	0.10
	rs4148330	G/A	17	22	4	35	0.40
ABCC2	rs2273697	G/A	32	19	2	22	0.68
	rs3740066	G/A	23	21	8	36	0.39
	rs717620	G/A	28	13	1	18	0.72
ABCC3	rs4793665	T/C	17	30	9	43	0.48

Gene	SNP	Alleles	N11 (%)	N12 (%)	N22 (%)	MAF (%)	HWE
	rs2107441	A/G	16	23	6	39	0.61
	rs2412333	G/A	23	24	4	31	0.50
	rs733392	G/A	21	28	3	33	0.10
	rs12602161	A/G	37	14	0	14	0.10
ABCC10	rs1214748	G/A	15	28	5	40	0.12
	rs831314	A/G	37	13	2	16	0.53
	rs1214752	G/A	19	18	8	38	0.31
ABCG2	rs2231142	C/A	39	13	0	13	0.30
GGH	rs3758149	C/T	20	28	7	38	0.56
SLC19A1	rs1051266	A/G	15	27	14	49	0.79
NR1I2 (SXR)	rs7643038	A/G	16	20	9	42	0.55
	rs3814055	G/A	18	23	11	43	0.47
	rs1054190	G/A	36	12	0	13	0.32
	rs3732361	G/A	13	24	9	46	0.72
	rs3814058	A/G	25	16	3	25	0.84
	rs6785049	A/G	11	29	9	48	0.19

5.7.1 Patient data

59 patients participated in the study, all of whom were diagnosed with osteosarcoma between 1988 and 2006 at the 2nd Department of Paediatrics of Semmelweis University. The participants are all Hungarian. Informed consent from patients or the parents was received, and the whole study was carried out according to the principles expressed in the Declaration of Helsinki, approved by the Hungarian Scientific and Research Ethics Committee of the Medical Research Council (case no.: 8-374/2009-1018EKU 914/PI/08.).

The clinical data contain 551 blocks of methotrexate treatments, with the dosage of 12g/m² body surface area, applied 4 to 12 times in every case. The patient database available for this study contains the following collected information: age at diagnosis; gender; risk-group; serum MTX level at 6h, 24h, 36h and 48h after treatment; lowest serum total protein; white blood cell count; neutrophil granulocyte count; highest value of alanine aminotransferase (ALAT) and aspartate aminotransferase (ASAT); bilirubin and creatinine during 2 week after treatment.

The following derived measures of the pharmacokinetics were used: area under the concentration–time curve in the first 48 hour (AUC₀₋₄₈), the peak methotrexate concentration, and the half-lives of methotrexate: T1 and T2 assuming two-compartment kinetics. T1 and T2 were derived using the serum level measurements before and after 24h respectively.

The main tool to analyse this database was univariate frequentist statistic (Pearson's chi-squared test), carried out and interpreted as a part of a parallel work [93]. We used the results of the Bayesian multilevel relevance analysis to complement and confirm the frequentist results in that work, specifically with respect to interactions. For the above mentioned reasons, here we will discuss details of the frequentist methodology only when it is essential for clear understanding.

5.7.2 Bayesian multilevel relevance analysis

We used Cooper-Herskovits (CH) non-informative structure prior for the analysis [94]. The maximal number of parents per node is a parameter of the method, which we set to 4 and to 2 in two separate runs. This setting limits the space of all possible Bayesian network models in case of limited data availability. We used 200 million Markov-chain

Monte Carlo (MCMC) steps; one million out of which were discarded from the beginning as burn-in to ensure that our samples are drawn from the correct *a posteriori* distribution. We only accepted associations, which appeared in the case of both parent count setting.

6 Results

6.1 Fusion of heterogeneous information sources for the prediction of the biological effect of small-molecular drugs

The aim of the research we conducted into computational drug repositioning was to compare the predictive performance of the newly developed Kernel Fusion Repositioning (KFR) method as an *intermediate* fusion method and a standard *late fusion* method, the Borda protocol based fusion via using one-class support vector machines as the model class. The Level 4 ATC classes were used as prediction tasks.

We computed AUC[ROC], AUC[CROC(exp)], BEDROC, TOP25 and TOP100 Sensitivity and Specificity values for all prediction tasks, here ATC classes, and illustrated the result on boxplots (See Figure 23). Because of the high specificity values, they are also shown on Figure 24, with an appropriate range.

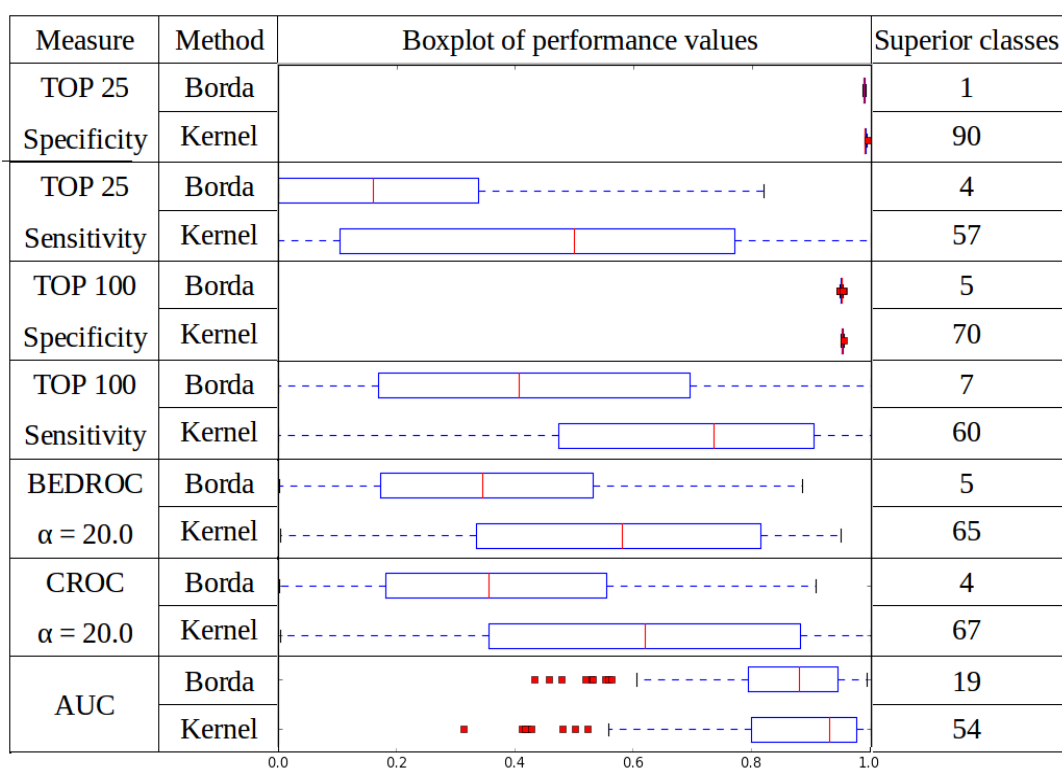






Figure 23 - Comparison of the performance of the intermediate and the late fusion method.

We also calculated the number of ATC classes which are significantly better predicted by the two fusion methods according to all measures (t-test; $p < 0.001$). We have found that in all cases, primarily underpinned by the early discovery measures, the intermediate data fusion has better predictive performance.

Measure	Method	Boxplot of performance values	Superior classes
TOP 25 Specificity	Borda		1
	Kernel		90
TOP 100 Specificity	Borda		5
	Kernel		70

0.90 0.92 0.94 0.96 0.98 1.00

Figure 24 - Comparison of specificities of the intermediate and the late fusion method.

To illustrate the result of the prioritization, a heatmap with hierarchical co-clustering is generated (see Figure 25). Every row in the heatmap corresponds to a drug and every column corresponds to a level 4 ATC class. The map is coloured according to the predicted membership relation between the drug and the class. Red signifies strongly predicted memberships, while blue signifies weak or no relations at all.

The hierarchical clustering organized the drugs with similar membership profiles, and the classes with similar members together, forming rectangular block structures in the map.

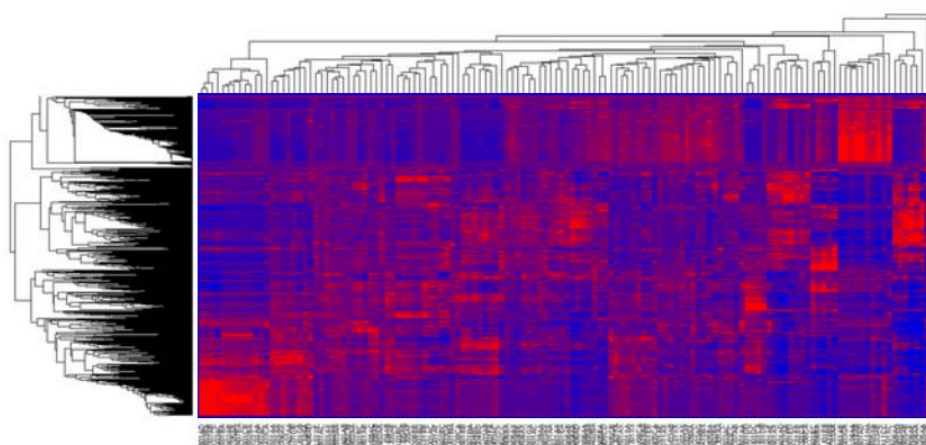


Figure 25 - Illustration of the drug-ATC class heatmap. Red colour signifies strong membership relations, blue signifies no membership relations.

An 8 x 16 (compound by ATC class) section of the heatmap in Figure 25 corresponds to some monoamine reuptake inhibitors shown on the Figure 26. All the drugs are either selective serotonin reuptake inhibitor (SSRI) (fluvoxamine, sertraline, paroxetine, fluoxetine, citalopram, escitalopram) or tricyclic antidepressants (protriptyline, nortriptyline). The red column shows the SSRI ATC class N06AB. There are antihistamine ATC classes (R06AD, R06AX, D04AA) in the neighbourhood, which can be a chemical structure related similarity, see e.g. fluoxetine and diphenhydramine. There are other classes like anti-obesity drugs (A08AA), erectile dysfunction related drugs (G04BE) or antiepileptics (N03AX) where the similarity can be anticipated based on biological knowledge. It is important to note that citalopram and escitalopram have slightly different profiles even with non-stereospecific chemical descriptors. This discrimination power comes from the other data sources.

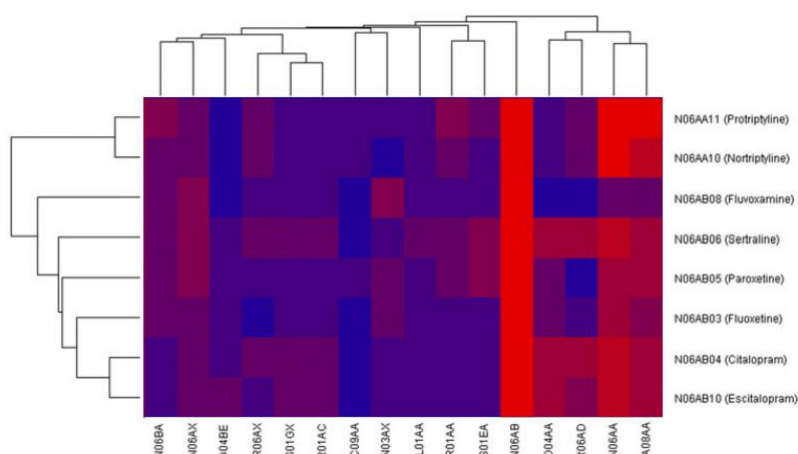


Figure 26 – Heatmap of monoamine reuptake inhibitor drugs and relevant ATC classes.

Some of the relevant classes are: non-selective monoamine reuptake inhibitors (N06AA), SSRIs (N06AB), other antidepressants (N06AX), sympathomimetic (N06BA, R01AA, S01EA), centrally acting antiobesity products (A08AA), erectile dysfunction related drugs (G04BE), antihistamines (R06AX, R06AD, D04AA) and antiepileptics (N03AX).

An additional direct output we can extract from the kernel fusion based technique is the weighting of the information sources (see Figure 27). The plot shows the average kernel weights of the 100 cross-validation runs.

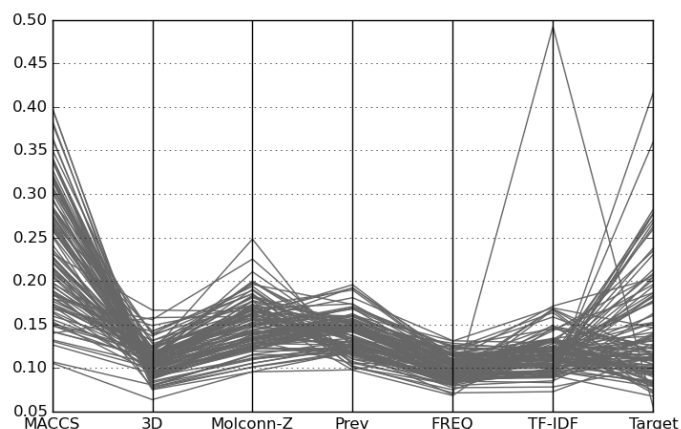


Figure 27 - Parallel coordinates diagram of the kernel weights: the relative importance of the different data sources determined by the KFR algorithm.

Since the Borda method does not have explicit weights, we calculated the Spearman correlation of the ordering based on the given single data source and the output orderings of the two fusion methods to compare their behaviour (see Figure 28). This measure is univariate, while the kernel weighting is multivariate in nature. This means that if two information sources are redundant, the kernel weights will drop, while the correlation between the output and the single source models will not.

A notable feature of these results, also a key result of my work, is that the relative contributions of the different data sources are quite stable across the different drug

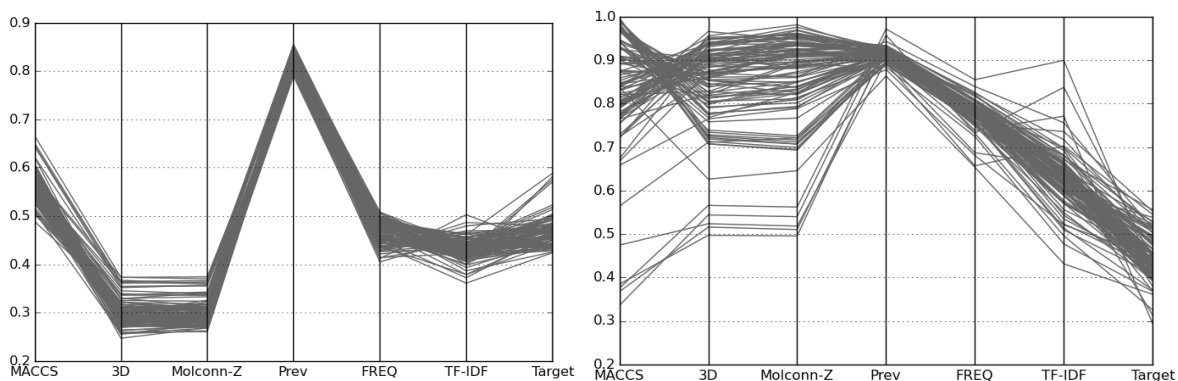


Figure 28 – Parallel coordinates diagrams of the Spearman correlations between the single source and the fusion models. The contributions are quite stable across the different drug categories in case of the Borda method (left), while the kernel fusion based method (right) shows adaptive, query-specific properties.

categories in case of the Borda method, while the kernel fusion based method shows adaptive, query-specific properties.

We observed cases, independently of the fusion method, where the predictive performance is less than $AUC = 0.5$, which means it is worse than the performance of a random model. These anomalous cases have to be removed to ensure applicability. The following solution forms a key result of my work: we suggested a criterion on query compactness to define an acceptable training set for prioritization [10]. The proposed solution relies on the use of the intraset similarity (ISS) to measure the diversity of a training set, where ISS is the average of all pairwise similarities of the elements in the training set T :

$$ISS = \sum_{i \in T} \sum_{j \in T, j \neq i} k_{i,j}$$

We normalized it with the average of all similarities in the full set of drugs: the universe of our experiment, called universal average similarity (UAS):

$$UAS = \sum_{i \in U} \sum_{j \in U, j \neq i} k_{i,j}$$

The measure ISS/UAS shows a good correlation with AUC values as it is shown on Figure 29. It can be seen that all classes which have higher than one ISS/UAS value, have at least 0.5 AUC.

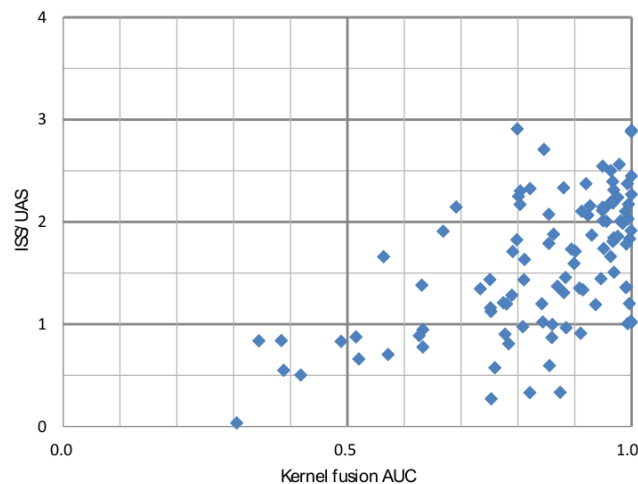


Figure 29 - The correlation of the ISS/UAS measure and the predictive performance.

In Table 12 and Table 13 the 10 most compact and the 10 least compact ATC classes are presented with their average pairwise similarity (ISS) values. It can be seen that the most compact ones are defined based on target or chemical class, while the diverse ones are based on broad functional categories.

Table 12 - The 10 most compact ATC Level 4 classes with the computed kernel-wise average ISS and ISS/UAS values.

ATC Level 4	Name	ISS	ISS/UAS
C07AB	Selective beta blocking agents	0.40852	2.90816
N02CC	Selective 5HT1 agonists	0.40679	2.89581
N06AB	Selective serotonin reuptake inhibitors	0.40474	2.88124
N05AB	Phenothiazines with piperazine structure	0.38035	2.70757
C09AA	Angiotensin-converting enzyme inhibitors	0.35983	2.5615
H02AB	Glucocorticoids	0.35725	2.54319
R06AA	Aminoalkyl ether antihistamines	0.35109	2.49929
L01DB	Anthracyclines and related substances	0.34375	2.44708
D07AB	Corticosteroids, moderately active (group II)	0.33612	2.39275
N04BC	Dopamine agonists	0.33329	2.37259

Table 13 - The 10 least compact ATC Level 4 classes with the computed kernel-wise average ISS and ISS/UAS values.

ATC Level 4	Name	ISS	ISS/UAS
A06AD	Osmotically acting laxatives	0.03782	0.26924
V08AC	Water soluble hepatotropic X-ray contrast media	0.04659	0.33164
G01AA	Gynecological antibiotics	0.04661	0.33178
V08CA	Paramagnetic contrast media	0.08047	0.57282
D06AX	Other antibiotics for topical use	0.08361	0.59510
S02AA	Otological antiinfectives	0.09246	0.65817
D01AE	Other antifungals for topical use	0.09879	0.70329
B05XA	Electrolyte solutions	0.10896	0.77562
D06BB	Antivirals for topical use	0.1135	0.80801
A07AA	Antibiotics, Intestinal	0.12187	0.48414

The geometry of this anomalous behaviour is illustrated on Figure 30. If the query is not compact, like the set of red dots on the figure, the model which separates them from the origin will rank a lot of unrelated compounds higher than the query itself (dots between the two groups on the figure). The compound ranked as 9th is more similar to the subgroup formed by the 7th and the 5th than the compounds ranked in the first place.

Both the MKL method and the single data source method applied in this comparison are sensitive to this situation, therefore it does not influence the comparison.

This behaviour, while presented here as anomalous, can be useful to detect outliers in a

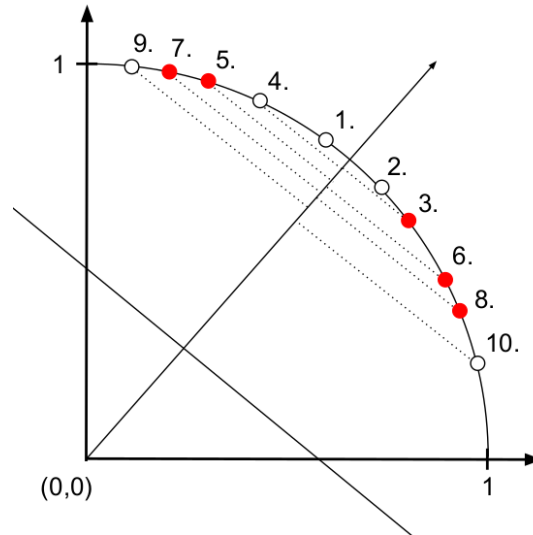


Figure 30 - Geometric illustration of the anomalous behaviour in the case of a heterogeneous query. The query compounds (red dots) are so heterogeneous that they are not ranked at the top of the list.

query, or what is the main goal here, to detect some novel entities with the same property as the query.

6.2 Application of the Kernel Fusion Repositioning method for finding Parkinson's disease related drugs

We analysed the result given to four Parkinson's disease (PD) related queries composed of neuroprotective agents, dopaminergic agents, muscarinic agents and NMDA antagonists by the KFR system (see Table 14) [14].

Table 14 - The four Parkinson's disease related queries

Query	Drugs
Neuroprotective agents	amantadine, pramipexole, rasagiline
Dopaminergic agents	bromocriptine, cabergoline, rotigotine, pramipexole
Muscarinic agents	biperiden, benztropine, trihexyphenidyl
NMDA antagonists	ifenprodil, budipine, amantadine, memantine

The neuroprotective query is the most heterogeneous one, containing compounds with different structural scaffolds and mechanisms of action, which can have an effect on the disease progression (See Figure 31).

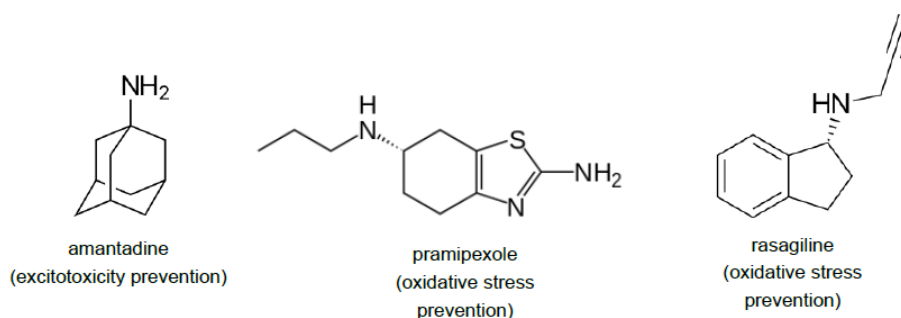


Figure 31 - Neuroprotective agents with their assumed neuroprotective mechanisms

Amantadine is an adamantane derivative originally introduced to the market as an antiviral agent inhibiting the M_2 protein of influenza A viruses [95]. It is a relatively weak

NMDA receptor antagonist and only indirectly increases dopamine release [89, 96]. Pramipexole is a dopamine agonist partially selective for D₃ receptor and an antioxidant [97, 98]. Both enantiomers of pramipexole can inhibit the mitochondrial production of reactive oxygen species (ROS) [98]. Rasagiline is an irreversible monoamine oxidase B (MAO-B) inhibitor, and a well-tolerated drug in PD therapy [99]. The effect against oxidative stress is only partly due to its MAO-B inhibitory effect [100, 101]: the reaction catalysed by MAO-B itself leads to H₂O₂ production, and in the next step to ROS production by Fenton's reaction. Another possible mechanism is a direct antioxidant effect due to the presence of the propargyl moiety [102].

The dopaminergic agonist query contains two ergoline (bromocriptine, cabergoline) and two non-ergoline (rotigotine, pramipexole) compounds (see Figure 32). Besides being a dopamine agonist, pramipexole is also have an antioxidant effect [97, 98].

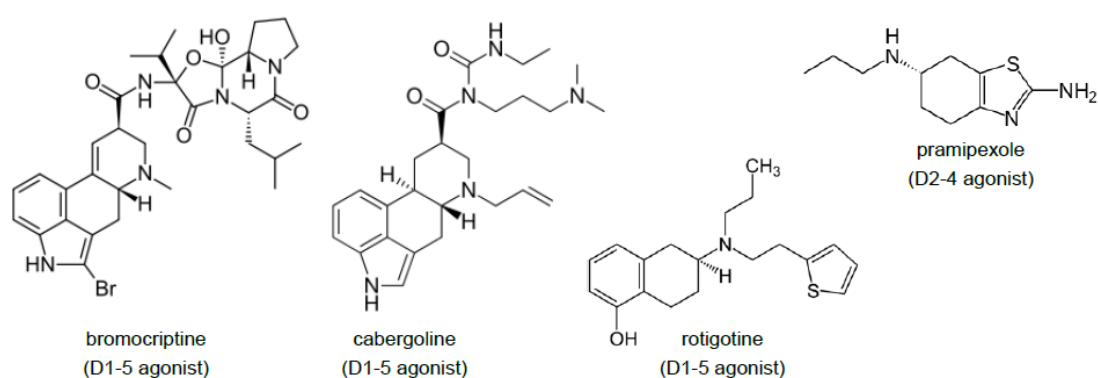


Figure 32 - Dopaminergic agonists and their main targeted subtypes [103].

The NMDA antagonist query is also structurally diverse containing two adamantane derivatives: amantadine, memantine (see Figure 33). In addition to its NMDA antagonist activity budipine shows anti-muscarinic effect as well [104]. Ifenprodil shows 400 fold selectivity for the NMDA receptor subunit NR2B relative to NR2A [105].

The members of the muscarinic antagonist group illustrated on Figure 34. Both biperiden and trihexyphenidyl show NMDA antagonist property as well [107].

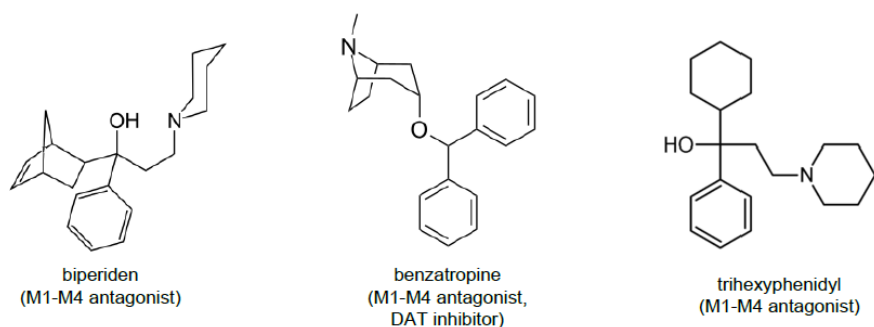


Figure 33 - NMDA antagonists with their specific features. Budipine shows anti-muscarinic effect [104], while ifenprodil shows selectivity based on NMDA receptor subunits [105, 106].

In Table 15 the result of the four prioritization runs is shown after the PubMed based filtering. For every query those top10 compounds are shown which have non-zero co-occurrence number defined with the following PubMed search query: („*Parkinson*” OR „*Parkinson's Disease*” OR „*PD*”) AND INN. As the original filtering was based on the state of the PubMed in 2013, the number of the found abstracts is also shown in case of a repeated search on the September 2016 version of the database. From a prospective point of view, which is the most reliable evaluation, it is interesting to note that the co-occurrence number for some of the highly prioritized compounds increased significantly. For example it is increased by 176% for clonidine and by 200% for gabapentin, while the relative increase was less significant for others (eg.: trihexyphenidyl or pergolide) or there was no change at all (eg.: encainide). These three groups show good correspondence with the following groups: possible repositioning candidates, already known drugs and false positives.

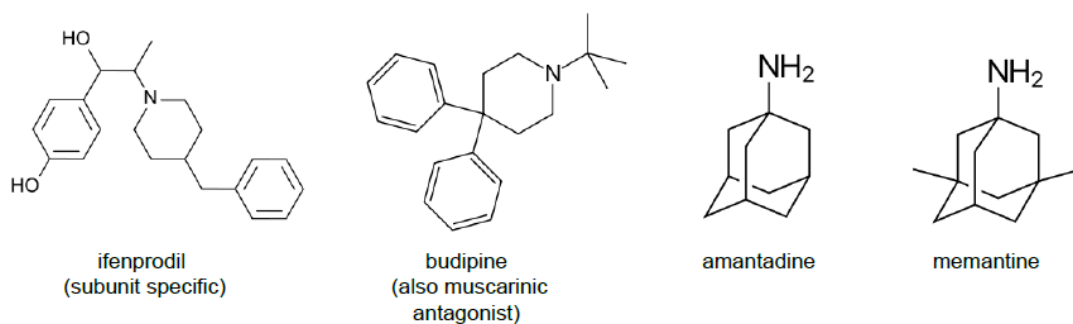


Figure 34 - Muscarinic antagonists with their targets [103].

Table 15 - Result of the prioritization with PD related queries. The prioritization list was filtered based on PubMed co-occurrence.

Query Description	Query elements with their resulted rank	Result	
		Ranking	PubMed hits 2013/2016
Neuroprotective agents	1 amantadine	5 memantine	77 / 166
	2 pramipexole	6 pergolide	442 / 550
	3 rasagiline	7 tacrine	30 / 45
		9 ropinirole	335 / 485
		12 gabapentin	20 / 60
		21 fentanyl	33 / 91
		24 ziprasidone	24 / 42
		25 clonidine	43 / 119
		26 chloroquine	11 / 55
		29 clozapine	289 / 446

Query Description	Query elements with their resulted rank	Result	
		Ranking	Query Description
Dopaminergic agonists	1 bromocriptine	Ranking	PubMed hits 2013/2016
	2 cabergoline		
	3 pramipexole	6 pergolide	442 / 550
	4 rotigotine	7 lisuride	240 / 272
		8 apomorphine	1105 / 1787
		9 risperidone	83 / 155
		10 aripiprazole	27 / 67
		11 ziprasidone	24 / 42
		13 olanzapine	92 / 161
		14 quetiapine	125 / 198
		15 ergotamine	12 / 16

Query Description	Query elements with their resulted rank	Result	
		Ranking	PubMed hits 2013/2016
NMDA antagonists	1 ifenprodil	8 dextromethorphan	19 / 46
	2 budipine	12 pergolide	442 / 550
	3 amantadine	13 aprindine	21 / 21
	4 memantine	16 benztropine	59 / 75
		19 mianserin	13 / 31
		20 imipramine	59 / 119
		21 biperiden	70 / 92
		23 encainide	25 / 25
		25 trihexyphenidyl	229 / 274
		29 donepezil	71 / 133

Query Description	Query elements with their resulted rank	Result	
		Ranking	Query Description
Muscarinic antagonists	1 biperiden	4 procyclidine	33 / 41
	2 benztropine	7 atropine	119 / 253
	3 trihexyphenidyl	21 dextromethorphan	19 / 46
		24 rotigotine	110 / 236
		32 perphenazine	37 / 54
		35 ajmaline	118 / 119
		38 quinidine	110 / 144
		40 haloperidol	335 / 665
		41 encainide	25 / 25
		42 donepezil	71 / 133

Beyond simply applying filters to lower the number of compounds we need to investigate, there are several other ways to extract information from the resulted ordering. One option is to use enrichment analysis to test if there is a property which is overrepresented in the top of the list. The application of enrichment analysis is discussed in our publication [12]. As an illustration, we show the application of CSEA, which can be seen as an extension of the prioritization method, which is also developed in our research group.

Using the information sources discussed in this work, and in addition a Connectivity Map based source described in our publication, we prioritized all compounds based on the similarity to amantadine [12]. We then calculated the enrichment of all ATC Level 4 classes in that list which is shown in Table 16. The original study is more detailed,

suggesting continuous information management through the drug discovery pipeline. However, since it is outside the scope of this thesis, here we refer to the original publication [12].

Table 16 – ATC Level 4 classes enriched in the list ordered by similarity to amantadine using all information sources + CMAP profiles. Detailed application scenario for CSEA is published in our paper [12].

Rank	ATC4	Name	E-value
1.	N04BC	Anti-Parkinson / Dopamine agonists	0.66352
2.	G03CC	Estrogens, combinations with other drugs	3.22836
3.	G02CB	Prolactine inhibitors	3.64457
4.	C03CA	Sulfonamides	4.10682
5.	C02CC	Guanidine derivative antihypertensives	5.93538
6.	N05AB	Phenothiazine antipsychotics with piperazine structure	7.72258
7.	A03FA	Propulsives	7.95207
8.	N05AE	Indole derivative antipsychotics	8.59378
9.	N07BB	Drugs used in alcohol dependence	11.0787
10.	N04AA	Anti-Parkinson / Tertiary amine anticholinergics	11.8485

As it is known that amantadine does not bind to the dopamine receptors, the presence of the class N04BC, or the classes G02CB, A03FA, N05AA which are rich in dopaminergic agents suggests indirect action on the dopaminergic system, which is well-known [108]. Other anti-Parkinson medications, like the ones in N04AA, also have an indirect effect on dopaminergic signalling [109].

6.3 Predicting multiple activities simultaneously improves the accuracy

Every level 4 ATC class contains a relatively low number of drugs, therefore learning a classifier which can generalize well is not easy. Some classes, however, show considerable similarity to one another. If two learning tasks are similar, we can use this similarity to learn them together, and this way we can increase the information available. To test this hypothesis we evaluated the predictive performance of Macau, described in Section 3.17, and compared it with a set of class by class trained regression models. As there is no negative set available, we used unscreened controls. For every positive sample in our dataset we randomly selected 4 membership relations from the unlabelled drug-class pairs and used them as negative set. We repeated this procedure 20 times, and using all the 20 datasets we trained models and averaged the predictions. The AUC values were computed for every ATC class using these aggregated predictions, and then these AUC values were averaged over the classes.

As Macau is a Bayesian method we do not need to set parameters to get the optimal performance. The only parameter we need to choose is the number of latent dimensions, but we know from our previous studies that choosing the latent dimension parameter slightly larger than necessary does not deteriorate the predictive performance [79, 80]. The correct strategy to choose the latent dimension parameter is to increase it as long as the performance increases, as larger value makes the algorithm slower without any gain.

The ridge regression has a regularization parameter λ , which we chose using a grid search, trying the values 0.01, 0.1, 1.0, 10.0, 100.0, 1000.0 and evaluated the performance using 30:70 class-level cross-validation.

Table 17 - Comparison of the average AUC of Macau and ridge regression. Macau is superior in all cases.

Information source	Macau		Ridge regression	
	Latent dims.	CV-AUC	Regularization	CV-AUC
MACCS	16	0.9135	10.0	0.8860
3D	16	0.9213	10.0	0.8226
MOLCONN-Z	12	0.9139	0.01	0.7316
TFIDF	16	0.8701	0.1	0.8559
TARGET	16	0.9146	0.1	0.8416

The result of the comparison is shown in Table 17. In all of the cases Macau has a considerably higher predictive performance than ridge regression. Presumably, Macau does not need as high-quality features as a single target method. One role of the feature in that case is to link compounds together and make the transfer of information between them possible.

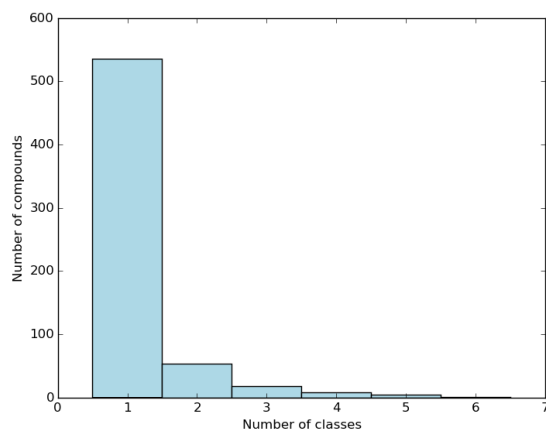


Figure 35 - Histogram of drugs involved in multiple classes. Most of the drugs involved only in one ATC level 4 class.

As the drug–ATC class matrix is very sparse, matrix factorization without side information cannot work on this dataset simply because most of the drugs are present only in one class (see Figure 35). This linking role can explain the relatively good performance of the MOLCONN-Z descriptor in the case of Macau, while it is a relatively poor predictor alone.

6.4 Comparison of BN-BMLA results to frequentist statistics in the task of associated variance detection for interpersonal methotrexate pharmacokinetics variability

As the clinical data contains 551 methotrexate blocks for 59 patients with a variable number of blocks per patient, the aggregation of the block level variables was necessary. I recommended using the median value over blocks for every variable as a patient level value, a convention used both in the frequentist and the Bayesian analysis. All pharmacokinetics and toxicity measurements were recorded at block level, therefore the median was computed.

To apply the BN-BMLA method the discretization of the continuous variables is necessary. I suggested the discretization based on median values to get a balanced dataset with an equal number of samples for different values. In case of multinomial variables we applied binning to binary variables to reach lower model complexity. The original hepatotoxicity and myelotoxicity variable was multinomial with four possible values. Based on the balanced dataset criterion the team binned the myelotoxicity as grade 1 vs. grade 2-4, and the hepatotoxicity as grade 1-2 vs. grade 3-4.

Both the toxicity variables (myelotoxicity, hepatotoxicity) and the pharmacokinetics parameters (AUC_{0-48} , peak methotrexate concentration, and methotrexate half-lives) show a strongly interconnected correlation structure with one another and with other clinical parameters. The BN-BMLA models show a connection between the pharmacokinetics (AUC_{0-48} , peak concentration) and the toxicity with *a posteriori* edge probability greater than 0.5. This connection was found based on frequentist methods as well [93]. There is also a strong link between the time of diagnosis (1988-1995 vs. 1996-2006) and the clinical parameters, which is due to the well-known fact of the different applied clinical protocol in these two periods [93].

In case of the NR1I2 gene two haplotype blocks can be identified. One of them is made up of two SNPs, the other is composed of three (see Table 18).

Table 18 - Haplotype blocks in the NR1I2 (SXR) gene. (D' : normalised linkage disequilibrium constant, LOD : log of the likelihood odds ratio, r^2 : correlation coefficient) [110]

SNP1	SNP2	D'	LOD	r²
Block 1				
rs7643038	rs3814055	1.0 [0.91 – 1.0]	19.95	0.957
Block 2				
rs3732361	rs3814058	1.0 [0.9 – 1.0]	17.47	0.916
rs3814058	rs6785049	1.0 [0.76 – 1.0]	6.66	0.437
rs6785049	rs3732361	1.0 [0.9 – 1.0]	17.47	0.916

The frequentist approach found all NR1I2 polymorphisms in the second block associated with heptato- and myelotoxicity, while the BN-BMLA identified a single one in the multiple target case. This SNP (rs3814058) is the same as the one where the frequentist p-value indicates the strongest interaction, which shows that the BN-BMLA methodology can distinguish between direct relations and transitive relations. It is important to note that every statement about a direct relevance can be interpreted only with the assumption that there is no other unmeasured variable which can change the chain of relevance relations. It is possible for example, that there is another polymorphism which has a real functional role, and even the identified rs3814058 SNP is only a marker, which is associated because it is in linkage disequilibrium with the functional polymorphism.

Table 19 - Effect size of identified SNPs. Effect size measured by the change of AUC and T_1 half-life, and by odds ratio (OR) in case of hepato- and myelotoxicity.

Frequentist	Bayesian	Gene	SNP	AUC	T_1	Hepato	Myelo
X		ABCC2	rs3740066	+2.8%	+2.0%	0.7778	0.2698
X		ABCG2	rs2231142	-16.3%	-8.3%	0.7742	0.3482
X		NR1I2	rs7643038	+16.2%	+18.9%	0.7653	0.5714
X			rs3814055	+7.3%	+16.6%	0.7653	0.5714
X			rs3732361	-5.7%	-1.8%	0.5625	0.9333
X			rs6785049	-1.5%	+1.5%	1.4624	1.5111
X	X		rs3814058	+2.2%	+7.9%	0.3333	1.6714
X	X		ABCB1	rs9282564	+16.0%	-0.9%	1.0345
X	X	ABCC3	rs4793665	+10.5%	+4.5%	1.0909	0.8182
X	X	ABCC2	rs717620	+26.5%	+16.2%	1.6800	0.2667
	X	ABCC1	rs246219	-6.2%	-2.3%	1.5714	2.6154
	X	GGH	rs3758149	+15.4%	+21.8%	0.5000	3.2500

While some SNPs are identified by both frameworks (two SNPs in case of AUC, and one in case of myelotoxicity), there are weaker candidates suggested by both the frequentist and the Bayesian methodology (see Table 19). Our results suggest that using the consensus of different methods for robust detection of association is an appropriate pragmatic approach to be followed.

7 Discussion

7.1 Fusion of heterogeneous information sources for the prediction of biological activity

The major goal of the research we conducted into drug repositioning was to compare the *late fusion* and the *intermediate fusion* paradigm via using the small molecule - ATC class membership prediction task as a gold standard. We found that the intermediate fusion shows better performance both in the case of unweighted AUC measure, and in the case of early discovery setting. In accordance with other observations in the literature we hypothesise that the difference is mainly due to the capability of the intermediate fusion to distinguish between within-source and between-source interactions [73], as it is beneficial to restrict the modelled correlation structure of the input space in case of high input dimensionality [74]. In the extreme case of dropping all interactions between features, we get the naive Bayes classifier [111], a well-known, simple and well-performing model in the case of high input dimensionality and a relatively small number of samples.

Our results showed that the model is capable of reconstructing meaningful membership relations, and deriving similarity between drugs and between classes. We illustrated this fact using the example of the SSRI and tricyclic antidepressant drugs on a co-clustered heatmap.

We witnessed that there is an optimal level of query heterogeneity. An appropriate level of heterogeneity can help us discover new results, but if the query is extremely heterogeneous, an anomalous behaviour takes place. In that case most of the candidate set is prioritized higher than the query. One of my key contributions was a diagnostic criterion, which can be used to filter these anomalous cases.

Our method is able to determine weights for the information sources simultaneously with the prioritization. The optimization of these data source weights with the primary goal of good predictive performance results in a very intuitive criterion. The optimal weights are those which make the query as compact as possible. For example, if a query is chemically compact, the chemical descriptors will get high weights. This adaptive weighting results

in a different level of incorporation of the data sources depending on the query. This property is an important advantage, which makes the method applicable in a wide range of pharmacological groups and different chemical spaces.

From a machine learning point of view, the problem discussed here is very similar to the problem of gene prioritization [112]. In case of gene prioritization the goal is to predict gene-disease associations, and we use a set of genes to represent a disease. Support vector machines and Multiple Kernel Learning were applied successfully in gene prioritization as well [113, 114].

7.2 Application of the Kernel Fusion Repositioning framework to find Parkinson's disease related drugs

The developed prioritization method was applied to search for Parkinson's disease related drugs, and it was able to identify other drugs used in the treatment of Parkinson's disease or co-occurring with the disease in the literature.

The prioritization for dopaminergic agonists clearly shows one of the limitations of our approach. We retrieved 5 antipsychotics in the top10 filtered list, which are well known antagonists at different dopaminergic receptors, particularly at D₂. Predicting target binding is evidently an easier task than predicting the functional role. This behaviour is expected even in case of the chemical structure based prediction, but it is more profound in case of the target data source. This source contains only the targets known for a given drug while the nature of the interaction on the targets are not encoded. The side effect data source can ameliorate this fallacy to some extent as side effects are consequences of functional effects. This phenomenon shows the importance of system level information sources like the side effect based profiles or gene expression profiles.

Our PubMed based filtering is clearly suboptimal as it removes totally new repositioning candidates while still leave false positives in the list. For example antipsychotics are not only used in the treatment of Parkinson's related psychosis, but they can also cause symptoms similar to Parkinson's disease. This filtering will therefore not eliminate the false positives generated by the fact that functional interaction prediction is difficult in our context. It is crucial to consider that a bad filtering can have a serious detrimental

effect on the predictions. However, if we use the PubMed co-occurrences in a prospective way, we can get much more credible signals.

Finally, we demonstrated the use of enrichment analysis tools in the interpretation of the Kernel Fusion Repositioning results. With enrichment analysis we showed that KFR is capable of retrieving different but related mechanism of actions to a query compound.

7.3 Prediction of multiple targets simultaneously

Our goal was to compare multiple target prediction with the classical single target prediction and we found that the predictive power is consequently higher for the multi-target method in case of each information source. Similar results were found in our previous pharmacogenomics studies predicting median inhibitory concentration (IC₅₀) values of compounds simultaneously on multiple targets [79].

Macau has a limitation in the area of fusion of multiple side information, as it can handle only fully observed side information. Side information matrices available for a different set of compounds cannot be concatenated. One option would be to drop all compounds which are not represented in all information sources, but it would result in a great waste of the available data. Other methods relying on the kernel trick have complementary application profile to Macau. While the latter is optimized for millions of matrix rows, here compounds, other methods are more suitable for a lower number of matrix rows but in exchange for a large number of features and several different information sources [115, 116]. On the other hand, we are working on making Macau capable of incorporating non-complete side information expressed in a non-kernelized form.

7.4 Advantages of Bayesian methods

The application of two novel Bayesian methods was discussed in the present work. In the following the common advantages of these methods will be discussed focusing on their theoretical relatedness.

We found direct probabilistic statements useful in both application areas. In the case of BN-BMLA the feedback of researchers in the field of genetics suggests that direct probabilistic statements are more natural than the frequentist viewpoint centred on the Type I (false positive) error. We have a similar experience with Macau: we got a specific

request to show a measure of credibility for our predictions to the pharmaceutical development team in an industrial scale drug-protein interaction prediction project. These questions are typically in the following form: “What is the probability that I will get a hit with $IC_{50} < 10\mu M$ if I test compound1 in my assay?”.

Furthermore, these methods offer some convenient advantages both in exploratory data analysis and in black-box modelling. In classical frequentist association studies we need to correct for multiple hypothesis testing as our main concern is the false positive type of error. The simplest solution is to use Bonferroni correction, which means we divide our significance level by the number of tests. This seriously reduces our statistical power because of the assumption of total independence between tests. We can apply more sophisticated correction mechanisms, but in case of multivariate Bayesian modelling we do not need to do so, as it is implicitly handled by the framework [117]. This implicit “correction” corresponds to the dependence structure of the variables and it is not more conservative than necessary.

In case of black box modelling, as all model parameters are treated as variables, their distributions are determined in the same framework as the prediction. Cross-validation does not need to be used to set the parameters. The only step to be handled is prior selection: we either use expert knowledge or select our priors to be non-informative.

Another significant advantage of Bayesian models is that they usually outperform their frequentist alternatives especially in the case of low sample sizes. It is observed that the Bayesian Probabilistic Matrix Factorization (BPMF) approach, one of the successors of Macau, outperforms the non-Bayesian matrix factorization especially in rows which are really sparse [118].

8 Conclusions

The results of my research allows for making the following conclusions and statements:

- While conducting my research I significantly contributed to and participated in the development process of a novel *intermediate* data fusion method, the Kernel Fusion Repositioning (KFR) framework. Our research evaluations showed that KFR has a superior performance compared to the *late fusion* baseline Borda protocol as justified by the AUC measure, and especially verified by all applied early discovery measures in terms of a drug repositioning benchmark problem.
- In order to examine the behaviour of the methods I analysed the Spearman's rank correlation of the single data source based prioritization results with the data fusion based prioritization results and we found that KFR shows adaptive, query-driven properties. This property is an important advantage, which makes the method applicable in a wide range of pharmacological groups and different chemical spaces.
- The experiments showed an anomalous behaviour in case of extremely high query heterogeneity and we witnessed that the query compounds are not ranked high in the resulted ordering. In this case the predictive power of the method can be really poor. We suggested a criterion measuring the average pairwise similarity of the query compounds to filter these cases, and showed that this criterion can identify the queries resulting in poor predictive performance.
- The KFR framework was applied to identify potential repositioning candidates in Parkinson's disease therapy and compounds showing high co-occurrence with those in the literature were retrieved. All results were validated further in a prospective evaluation. Also, steps of a novel computational route for drug repositioning candidate identification were outlined.
- I participated in the development process of Macau, a novel Bayesian matrix factorization method capable of predicting multiple targets simultaneously. While

conducting the research I compared Macau to a single target method (Ridge regression) and found it superior in the case of all information sources.

- My research justifies successful adaptation and application of BN-BMLA, a novel multivariate Bayesian method, in complementing and confirming the already existing frequentist results in a study conducted into the pharmacokinetics of high dose methotrexate therapy. The results suggest that the effective combination of the Bayesian and frequentist methods in the field of the robust detection of association is an appropriate strategy, whereas the BN-BMLA method is more beneficial in the case of investigating interactions or redundancies, such as linked polymorphism.

9 Summary

As the research and development productivity decreases, the pharmaceutical industry is continuously searching for new approaches in drug discovery to keep their business operational. Two possible options discussed in my work are drug repositioning and personalized medicine. In the age of big data, shared databases and precompetition time collaboration; information technologies, statistics and machine learning play an important role in these fields.

I significantly contributed to an interdisciplinary project in which we designed and implemented a data fusion method called Kernel Fusion Repositioning (KFR). KFR can predict the biological effects of small-molecular drugs using a diverse set of heterogeneous information sources. In my doctoral research I demonstrated that the kernel fusion framework shows better predictive performance than the early data fusion. The results show that there is an optimal level of heterogeneity of the query to discover new indications without getting anomalous behaviour.

The data fusion method was applied in order to identify Parkinson's disease related drugs. We observed that the method is capable of retrieving other drugs used in the clinical practice or drugs co-occurring in the literature with Parkinson's disease. Also, steps of a novel computational route for drug repositioning candidate identification were outlined.

I participated in the development of Macau, a novel Bayesian matrix factorization method capable of predicting multiple targets simultaneously. While conducting the research I compared Macau to a single target baseline and found it superior in the case of all information sources.

In addition to drug repositioning I also participated in a research project conducted into the pharmacokinetics of methotrexate at high dose levels. I adapted and applied a novel Bayesian multivariate statistical technique to identify predictive genetic variants for the interpersonal variability of methotrexate pharmacokinetics. Polymorphisms significantly overlapping with those independently discovered by frequentist methods were successfully retrieved, and the advantages of the new method were verified in case of linked polymorphisms and multiple target variables.

10 Összefoglalás

Ahogy a kutatás-fejlesztés hatékonysága csökken, a gyógyszeripari vállalatok a gyógyszerfejlesztés új irányaira kényszerülnek, hogy továbbra is releváns piaci szereplők maradjanak. A dolgozatomban tárgyalt két lehetséges út a gyógyszer újrapozicionálás és a személyre szabott gyógyászat. A megosztott adatbázisok és a korai fázisú gyógyszeripari együttműködések korszakában nagy szerep jut az információtechnológia és a gépi tanulás módszereinek.

Jelentős szerepet töltöttem be egy adatfúziós módszer, a Kernel Fusion Repositioning (KFR) keretrendszer megtervezését és implementálását célzó interdiszciplináris kutatásban. A KFR rendszer alkalmas kismolekulás vegyületek biológiai hatásának előrejelzésére heterogén információforrások felhasználásával. A doktori munkám során megmutattam, hogy a kernel fúziós keretrendszer előrejelzési pontossága felülmúlja az úgynevezett korai adatfúziós megközelítés eredményeit. Az eredmények tükrében kijelenthető továbbá, hogy létezik a lekérdezési gyógyszerhalmaznak egy optimális heterogenitása, amely mellett feltárhatók új indikációk ugyanakkor elkerülhető a módszer rendellenes működése.

Ezt követően Parkinson-kór kezelése szempontjából releváns gyógyszerjelöltek keresésére alkalmaztam a fenti adatfúziós eljárást, és megfigyeltem, hogy a módszer alkalmas klinikai gyakorlatban alkalmazott gyógyszerek és az indikációt tekintve új, a szakirodalomban a Parkinson-kórral együttesen előforduló vegyületek megtalálására. Továbbá vázoltam egy számítógépes módszereket használó újszerű munkafolyamat lépéseit, mely alkalmas újrapozicionálási jelöltek azonosítására.

Részt vettem egy több célváltozó együttes becslésére képes mátrix faktorizációs módszer, a Macau kifejlesztésében. Jelen kutatás keretében összehasonlítottam a Macau-t egy egyváltozós módszerrel, és a pontosabbnak találtam a használt információforrástól függetlenül.

A fentiekén túl részt vettem egy kutatásban, amely a nagy dózisban adagolt metotrexát farmakokinetikáját vizsgálta. Adaptáltam és alkalmaztam egy új Bayes-i többváltozós statisztikai technikát a metotrexát farmakokinetika betegenkénti variabilitásának szempontjából prediktív genetikai variánsok azonosítására. Az általam azonosított

polimorfizmusok jelentős átfedést mutattak a frekventista módszerek használatával azonosítottakkal. Ezen felül megmutattam az új módszer előnyeit kapcsolt polimorfizmusok és több célváltozó együttes vizsgálata esetén.

References

1. Kinch MS, Haynesworth A, Kinch SL, Hoyer D. (2014) An overview of FDA-approved new molecular entities: 1827-2013. *Drug Discov Today*, 19: 1033-9.
2. Pammolli F, Magazzini L, Riccaboni M. (2011) The productivity crisis in pharmaceutical R&D. *Nat Rev Drug Discov*, 10: 428-38.
3. Moors EH, Cohen AF, Schellekens H. (2014) Towards a sustainable system of drug development. *Drug Discov Today*, 19: 1711-20.
4. Ashburn TT, Thor KB. (2004) Drug repositioning: identifying and developing new uses for existing drugs. *Nat Rev Drug Discov*, 3: 673-83.
5. Mossinghoff GJ. (1999) Overview of the Hatch-Waxman Act and its impact on the drug development process. *Food Drug Law J*, 54: 187-94.
6. Hauben M, Madigan D, Gerrits CM, Walsh L, Van Puijenbroek EP. (2005) The role of data mining in pharmacovigilance. *Expert Opin Drug Saf*, 4: 929-48.
7. Arany A. Computational aspects of pharmaceutical research. In P. Antal (*edit.*), *Bioinformatics*. Typotex Kft., Budapest, 2014: 241-252.
8. Arrowsmith J, Miller P. (2013) Trial watch: phase II and phase III attrition rates 2011-2012. *Nat Rev Drug Discov*, 12: 569.
9. Lendrem D, Senn SJ, Lendrem BC, Isaacs JD. (2015) R&D productivity rides again? *Pharm Stat*, 14: 1-3.
10. Arany A, Bolgar B, Balogh B, Antal P, Matyus P. (2013) Multi-aspect candidates for repositioning: data fusion methods using heterogeneous information sources. *Curr Med Chem*, 20: 95-107.
11. Norton PA, Zinner NR, Yalcin I, Bump RC, Duloxetine Urinary Incontinence Study Group. (2002) Duloxetine versus placebo in the treatment of stress urinary incontinence. *Am J Obstet Gynecol*, 187: 40-8.
12. Temesi G, Bolgar B, Arany A, Szalai C, Antal P, Matyus P. (2014) Early repositioning through compound set enrichment analysis: a knowledge-recycling strategy. *Future Med Chem*, 6: 563-75.
13. Novac N. (2013) Challenges and opportunities of drug repositioning. *Trends Pharmacol Sci*, 34: 267-72.
14. Bolgar B, Arany A, Temesi G, Balogh B, Antal P, Matyus P. (2013) Drug repositioning for treatment of movement disorders: from serendipity to rational discovery strategies. *Curr Top Med Chem*, 13: 2337-63.
15. Li YY, Jones SJ. (2012) Drug repositioning for personalized medicine. *Genome Med*, 4: 27.
16. Dunkel P, Chai CL, Sperlagh B, Huleatt PB, Matyus P. (2012) Clinical utility of neuroprotective agents in neurodegenerative diseases: current status of drug development for Alzheimer's, Parkinson's and Huntington's diseases, and amyotrophic lateral sclerosis. *Expert Opin Investig Drugs*, 21: 1267-308.
17. Borish L, Culp JA. (2008) Asthma: a syndrome composed of heterogeneous diseases. *Ann Allergy Asthma Immunol*, 101: 1-8; quiz 8-11, 50.

18. Eckert H, Bajorath J. (2007) Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discov Today*, 12: 225-33.
19. Kubinyi H. (1998) Similarity and Dissimilarity: A Medicinal Chemist's View. *Perspectives in Drug Discovery and Design*, 9: 225-252.
20. Maggiora G, Vogt M, Stumpfe D, Bajorath J. (2014) Molecular similarity in medicinal chemistry. *J Med Chem*, 57: 3186-204.
21. Wolpert DH. (1996) The Lack of A Priori Distinctions Between Learning Algorithms. *Neural Computation*, 8: 1341-1390.
22. Amaratunga D, Cabrera J. (2001) Analysis of Data From Viral DNA Microchips. *Journal of the American Statistical Association*, 96: 1161-1170.
23. Bolstad BM, Irizarry RA, Astrand M, Speed TP. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19: 185-93.
24. Ginn CMR, Willett P, Bradshaw J. (2000) Combination of molecular similarity measures using data fusion. *Perspectives in Drug Discovery and Design*, 20: 1-16.
25. Willett P. (2006) Enhancing the Effectiveness of Ligand-Based Virtual Screening Using Data Fusion. *QSAR & Combinatorial Science*, 25: 1143-1152.
26. Whittle M, Gillet VJ, Willett P, Alex A, Loesel J. (2004) Enhancing the Effectiveness of Virtual Screening by Fusing Nearest Neighbor Lists: A Comparison of Similarity Coefficients. *Journal of Chemical Information and Computer Sciences*, 44: 1840-1848.
27. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A. (2004) Comparison of Fingerprint-Based Methods for Virtual Screening Using Multiple Bioactive Reference Structures. *Journal of Chemical Information and Computer Sciences*, 44: 1177-1185.
28. Zhang Q, Muegge I. (2006) Scaffold hopping through virtual screening using 2D and 3D similarity descriptors: ranking, voting, and consensus scoring. *J Med Chem*, 49: 1536-48.
29. Feher M. (2006) Consensus scoring for protein-ligand interactions. *Drug Discov Today*, 11: 421-8.
30. Plewczynski D, Spieser SA, Koch U. (2006) Assessing different classification methods for virtual screening. *J Chem Inf Model*, 46: 1098-106.
31. Terp GE, Johansen BN, Christensen IT, Jørgensen FS. (2001) A New Concept for Multidimensional Selection of Ligand Conformations (MultiSelect) and Multidimensional Scoring (MultiScore) of Protein–Ligand Binding Affinities. *Journal of Medicinal Chemistry*, 44: 2333-2343.
32. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A. (2006) New methods for ligand-based virtual screening: use of data fusion and machine learning to enhance the effectiveness of similarity searching. *J Chem Inf Model*, 46: 462-70.
33. Triguero I, García S, Herrera F. (2015) Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. *Knowledge and Information Systems*, 42: 245-284.
34. Yang JM, Chen YF, Shen TW, Kristal BS, Hsu DF. (2005) Consensus scoring criteria for improving enrichment in virtual screening. *J Chem Inf Model*, 45: 1134-46.

35. Svensson F, Karlen A, Skold C. (2012) Virtual screening data fusion using both structure- and ligand-based methods. *J Chem Inf Model*, 52: 225-32.
36. Hopkins AL. (2008) Network pharmacology: the next paradigm in drug discovery. *Nat Chem Biol*, 4: 682-90.
37. Csermely P, Agoston V, Pongor S. (2005) The efficiency of multi-target drugs: the network approach might help drug design. *Trends Pharmacol Sci*, 26: 178-82.
38. Roth BL, Sheffler DJ, Kroeze WK. (2004) Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nat Rev Drug Discov*, 3: 353-9.
39. Nijman SM. (2011) Synthetic lethality: general principles, utility and detection using genetic screens in human cells. *FEBS Lett*, 585: 1-6.
40. Barabasi AL, Gulbahce N, Loscalzo J. (2011) Network medicine: a network-based approach to human disease. *Nat Rev Genet*, 12: 56-68.
41. Wermuth CG. (2006) Selective optimization of side activities: the SOSA approach. *Drug Discov Today*, 11: 160-4.
42. Lamb J. (2007) The Connectivity Map: a new tool for biomedical research. *Nat Rev Cancer*, 7: 54-60.
43. Laenen G, Thorrez L, Bornigen D, Moreau Y. (2013) Finding the targets of a drug by integration of gene expression data with a protein interaction network. *Mol Biosyst*, 9: 1676-85.
44. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, Reich M, Hieronymus H, Wei G, Armstrong SA, Haggarty SJ, Clemons PA, Wei R, Carr SA, Lander ES, Golub TR. (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313: 1929-35.
45. Hu G, Agarwal P. (2009) Human disease-drug network based on genomic expression profiles. *PLoS One*, 4: e6536.
46. Laenen G, Ardeshirdavani A, Moreau Y, Thorrez L. (2015) Galahad: a web server for drug effect analysis from gene expression. *Nucleic Acids Res*, 43: W208-12.
47. Chiang AP, Butte AJ. (2009) Systematic evaluation of drug-disease relationships to identify leads for novel drug uses. *Clin Pharmacol Ther*, 86: 507-10.
48. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK. (2007) Relating protein pharmacology by ligand chemistry. *Nat Biotechnol*, 25: 197-206.
49. Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, Jensen NH, Kuijter MB, Matos RC, Tran TB, Whaley R, Glennon RA, Hert J, Thomas KL, Edwards DD, Shoichet BK, Roth BL. (2009) Predicting new molecular targets for known drugs. *Nature*, 462: 175-81.
50. Gottlieb A, Stein GY, Ruppin E, Sharan R. (2011) PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol Syst Biol*, 7: 496.
51. Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P. (2010) A side effect resource to capture phenotypic effects of drugs. *Mol Syst Biol*, 6: 343.
52. Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P. (2008) Drug target identification using side-effect similarity. *Science*, 321: 263-6.

53. Kuhn MC, M.; Bork,P.; Jensen,L.J.; Gavin,A.; Costi,M.P.; Luciani,R.; Preissner,R.; Fan,H.; Hossbach,J. Use of Aprepitant and Derivatives Thereof for the Treatment of Cancer. 2008: (*Patent*)
54. von Eichborn J, Murgueitio MS, Dunkel M, Koerner S, Bourne PE, Preissner R. (2011) PROMISCUOUS: a database for network-based drug-repositioning. *Nucleic Acids Res*, 39: D1060-6.
55. Truchon JF, Bayly CI. (2007) Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *J Chem Inf Model*, 47: 488-508.
56. Hanley JA, McNeil BJ. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143: 29-36.
57. Swamidass SJ, Azencott CA, Daily K, Baldi P. (2010) A CROC stronger than ROC: measuring, visualizing and optimizing early retrieval. *Bioinformatics*, 26: 1348-56.
58. Efron B. (2013) Mathematics. Bayes' theorem in the 21st century. *Science*, 340: 1177-8.
59. Cheng Y, Prusoff WH. (1973) Relationship between the inhibition constant (K_i) and the concentration of inhibitor which causes 50 per cent inhibition (I₅₀) of an enzymatic reaction. *Biochem Pharmacol*, 22: 3099-108.
60. Hegyi M, Arany A, Semsei AF, Csordas K, Eipel O, Gezsi A, Kutszegi N, Csoka M, Muller J, Erdelyi DJ, Antal P, Szalai C, Kovacs GT. (2016) Pharmacogenetic analysis of high-dose methotrexate treatment in children with osteosarcoma. *Oncotarget*, 8: 9388-9398.
61. Gezsi A, Lautner-Csorba O, Erdelyi DJ, Hullam G, Antal P, Semsei AF, Kutszegi N, Hegyi M, Csordas K, Kovacs G, Szalai C. (2015) In interaction with gender a common CYP3A4 polymorphism may influence the survival rate of chemotherapy for childhood acute lymphoblastic leukemia. *Pharmacogenomics J*, 15: 241-7.
62. Lautner-Csorba O, Gezsi A, Erdelyi DJ, Hullam G, Antal P, Semsei AF, Kutszegi N, Kovacs G, Falus A, Szalai C. (2013) Roles of genetic polymorphisms in the folate pathway in childhood acute lymphoblastic leukemia evaluated by Bayesian relevance and effect size analysis. *PLoS One*, 8: e69843.
63. Lautner-Csorba O, Gezsi A, Semsei AF, Antal P, Erdelyi DJ, Schermann G, Kutszegi N, Csordas K, Hegyi M, Kovacs G, Falus A, Szalai C. (2012) Candidate gene association study in pediatric acute lymphoblastic leukemia evaluated by Bayesian network based Bayesian multilevel analysis of relevance. *BMC Med Genomics*, 5: 42.
64. Varga G, Szekely A, Antal P, Sarkozy P, Nemoda Z, Demetrovics Z, Sasvari-Szekely M. (2012) Additive effects of serotonergic and dopaminergic polymorphisms on trait impulsivity. *Am J Med Genet B Neuropsychiatr Genet*, 159B: 281-8.
65. Antal P, Hullam G, Hajos G, Sarkozy P, Gezsi A, Szalai C, Falus A. Bayesian, Systems-based, Multilevel Analysis of Associations for Complex Phenotypes: from Interpretation to Decision. In R. Mourad (*edit.*), *Probabilistic Graphical Models for Genetics, Genomics, and Postgenomics*. Oxford University Press, Oxford, UK., 2014: 318-360.
66. Wold S, Ruhe A, Wold H, Dunn III W. (1984) The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, 5: 735-743.

67. Varmuza K, Filzmoser P. Introduction to multivariate statistical analysis in chemometrics. CRC Press, Boca Raton, 2009.
68. Caruana R. Algorithms and applications for multitask learning. in ICML. 1996.
69. Andersson M. (2009) A comparison of nine PLS1 algorithms. *Journal of Chemometrics*, 23: 518-529.
70. Cristianini N, Shawe-Taylor J. An introduction to support vector machines : and other kernel-based learning methods. Cambridge University Press, Cambridge ; New York, 2000: 22-24.
71. Golub GH, Van Loan CF. Matrix computations. Johns Hopkins University Press, Baltimore, 1996: 51.
72. Cristianini N, Shawe-Taylor J. An introduction to support vector machines : and other kernel-based learning methods. Cambridge University Press, Cambridge ; New York, 2000: 26-52.
73. Pavlidis P, Weston J, Cai J, Noble WS. (2002) Learning gene functional classifications from multiple data types. *J Comput Biol*, 9: 401-11.
74. Srivastava N, Hinton GE, Krizhevsky A, Sutskever I, Salakhutdinov R. (2014) Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15: 1929-1958.
75. Sun Z, Ampornpant N, Varma M, Vishwanathan S. Multiple kernel learning and the SMO algorithm. in *Advances in neural information processing systems*. 2010.
76. Vapnik VN. The nature of statistical learning theory. Springer, New York, 2000: 138-146.
77. Denis F, Gilleron R, Tommasi M. Text classification from positive and unlabeled examples. in *Proceedings of the 9th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU'02*. 2002.
78. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning : data mining, inference, and prediction. Springer, New York, NY, 2009: 241-249.
79. Simm J, Arany A, Zakeri P, Haber T, Wegner JK, Chupakhin V, Ceulemans H, Moreau Y. (2015) Macau: Scalable Bayesian Multi-relational Factorization with Side Information using MCMC. arXiv preprint arXiv:1509.04610.
80. Arany A, Simm J, Zakeri P, Haber T, Wegner JK, Chupakhin V, Ceulemans H, Moreau Y. (2015) Highly Scalable Tensor Factorization for Prediction of Drug-Protein Interaction Type. arXiv preprint arXiv:1512.00315.
81. Henderson HV, Searle SR. (1981) The vec-permutation matrix, the vec operator and Kronecker products: A review. *Linear and multilinear algebra*, 9: 271-288.
82. Bishop CM. Pattern recognition and machine learning. Springer, New York, 2006: 542 - 546.
83. Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, Djoumbou Y, Eisner R, Guo AC, Wishart DS. (2011) DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res*, 39: D1035-41.
84. NIH US. *DailyMed*. 2010; Available from: <https://dailymed.nlm.nih.gov/>.
85. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use, *Medical Dictionary for Regulatory Activities*. 2010.

86. Bodenreider O. (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res*, 32: D267-70.
87. Lamb J, Golub TR, Subramanian A, Peck DD. Gene-expression profiling with reduced numbers of transcript measurements. M.I.o. Technology 2014: WO2011US31395 (*Patent*)
88. Wawer MJ, Li K, Gustafsdottir SM, Ljosa V, Bodycombe NE, Marton MA, Sokolnicki KL, Bray MA, Kemp MM, Winchester E, Taylor B, Grant GB, Hon CS, Duvall JR, Wilson JA, Bittker JA, Dancik V, Narayan R, Subramanian A, Winckler W, Golub TR, Carpenter AE, Shamji AF, Schreiber SL, Clemons PA. (2014) Toward performance-diverse small-molecule libraries for cell-based phenotypic screening using multiplexed high-dimensional profiling. *Proc Natl Acad Sci U S A*, 111: 10911-6.
89. Lange KW, Riederer P. (1994) Glutamatergic drugs in Parkinson's disease. *Life Sci*, 55: 2067-75.
90. Haider N. (2010) Functionality pattern matching as an efficient complementary structure/reaction search tool: an open-source approach. *Molecules*, 15: 5079-92.
91. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102: 15545-50.
92. Stojmirovic A, Yu YK. (2010) Robust and accurate data enrichment statistics via distribution function of sum of weights. *Bioinformatics*, 26: 2752-9.
93. Hegyi M, *Analysis of prognostic factors, pharmacokinetic and pharmacogenetic examinations in children with osteosarcoma*, in *Doctoral School of Clinical Medicine*. 2013, Semmelweis University: Budapest.
94. Cooper GF, Herskovits E. (1992) A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9: 309-347.
95. Hay AJ, Wolstenholme AJ, Skehel JJ, Smith MH. (1985) The molecular basis of the specific anti-influenza action of amantadine. *EMBO J*, 4: 3021-4.
96. Kornhuber J, Weller M, Schoppmeyer K, Riederer P. (1994) Amantadine and memantine are NMDA receptor antagonists with neuroprotective properties. *J Neural Transm Suppl*, 43: 91-104.
97. Le WD, Jankovic J, Xie W, Appel SH. (2000) Antioxidant property of pramipexole independent of dopamine receptor activation in neuroprotection. *J Neural Transm (Vienna)*, 107: 1165-73.
98. Ferrari-Toninelli G, Maccarinelli G, Uberti D, Buerger E, Memo M. (2010) Mitochondria-targeted antioxidant effects of S(-) and R(+) pramipexole. *BMC Pharmacol*, 10: 2.
99. Reichmann H, Jost WH. (2010) Efficacy and tolerability of rasagiline in daily clinical use--a post-marketing observational study in patients with Parkinson's disease. *Eur J Neurol*, 17: 1164-71.
100. Pizzinat N, Copin N, Vindis C, Parini A, Cambon C. (1999) Reactive oxygen species production by monoamine oxidases in intact cells. *Naunyn Schmiedebergs Arch Pharmacol*, 359: 428-31.
101. Blandini F, Armentero MT, Fancellu R, Blaugrund E, Nappi G. (2004) Neuroprotective effect of rasagiline in a rodent model of Parkinson's disease. *Exp Neurol*, 187: 455-9.

102. Mandel S, Weinreb O, Amit T, Youdim MB. (2005) Mechanism of neuroprotective action of the anti-Parkinson drug rasagiline and its derivatives. *Brain Res Brain Res Rev*, 48: 379-87.
103. Roth BL, Lopez E. *Ki Database - Psychiatric Drug Screening Program*. 2000 [cited 2016; Available from: <http://kidbdev.med.unc.edu/databases/kidb.php>.
104. Klockgether T, Wullner U, Steinbach JP, Petersen V, Turski L, Loschmann PA. (1996) Effects of the antiparkinsonian drug budipine on central neurotransmitter systems. *Eur J Pharmacol*, 301: 67-73.
105. Williams K. (1993) Ifenprodil discriminates subtypes of the N-methyl-D-aspartate receptor: selectivity and mechanisms at recombinant heteromeric receptors. *Molecular pharmacology*, 44: 851-859.
106. Scheller D, Ullmer C, Berkels R, Gwarek M, Lubbert H. (2009) The in vitro receptor profile of rotigotine: a new agent for the treatment of Parkinson's disease. *Naunyn Schmiedebergs Arch Pharmacol*, 379: 73-86.
107. McDonough JH, Jr., Shih TM. (1995) A study of the N-methyl-D-aspartate antagonistic properties of anticholinergic drugs. *Pharmacol Biochem Behav*, 51: 249-53.
108. Breier A, Adler CM, Weisenfeld N, Su TP, Elman I, Picken L, Malhotra AK, Pickar D. (1998) Effects of NMDA antagonism on striatal dopamine release in healthy subjects: application of a novel PET approach. *Synapse*, 29: 142-7.
109. Dewey SL, Smith GS, Logan J, Brodie JD, Simkowitz P, MacGregor RR, Fowler JS, Volkow ND, Wolf AP. (1993) Effects of central cholinergic blockade on striatal dopamine release measured with positron emission tomography in normal human subjects. *Proc Natl Acad Sci U S A*, 90: 11816-20.
110. Barrett JC, Fry B, Maller J, Daly MJ. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, 21: 263-5.
111. Jordan A. (2002) On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 14: 841.
112. Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, Tranchevent LC, De Moor B, Marynen P, Hassan B, Carmeliet P, Moreau Y. (2006) Gene prioritization through genomic data fusion. *Nat Biotechnol*, 24: 537-44.
113. Yu S, Tranchevent LC, De Moor B, Moreau Y. (2010) Gene prioritization and clustering by multi-view text mining. *BMC Bioinformatics*, 11: 28.
114. Zakeri P, Elshal S, Moreau Y. Gene prioritization through geometric-inspired kernel data fusion. in *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*. 2015. IEEE.
115. Gönen M, Khan SA, Kaski S. Kernelized Bayesian Matrix Factorization. in *ICML (3)*. 2013.
116. Bolgár B, Antal P. Bayesian Matrix Factorization with Non-Random Missing Data using Informative Gaussian Process Priors and Soft Evidences. in *Proceedings of the Eighth International Conference on Probabilistic Graphical Models*. 2016.
117. Gelman A, Hill J, Yajima M. (2012) Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, 5: 189-211.

118. Salakhutdinov R, Mnih A. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. in Proceedings of the 25th international conference on Machine learning. 2008. ACM.

List of own publications

Hegy M, Arany A, Semsei AF, Csordas K, Eipel O, Gezsi A, Kutszegi N, Csoka M, Muller J, Erdelyi DJ, Antal P, Szalai C, and Kovacs GT, *Pharmacogenetic analysis of high-dose methotrexate treatment in children with osteosarcoma*. *Oncotarget*, 2016. **IF = 5.008***

Temesi G, Bolgar B, Arany A, Szalai C, Antal P, and Matyus P, *Early repositioning through compound set enrichment analysis: a knowledge-recycling strategy*. *Future Med Chem*, 2014. **6(5): p. 563-75. IF = 3.744**

Arany A, Bolgar B, Balogh B, Antal P, and Matyus P, *Multi-aspect candidates for repositioning: data fusion methods using heterogeneous information sources*. *Curr Med Chem*, 2013. **20(1): p. 95-107. IF = 3.715**

Bolgar B, Arany A, Temesi G, Balogh B, Antal P, and Matyus P, *Drug repositioning for treatment of movement disorders: from serendipity to rational discovery strategies*. *Curr Top Med Chem*, 2013. **13(18): p. 2337-63. IF = 3.453**

Acknowledgements

I would like to thank all of the help of my supervisor Prof. Péter Matyus, the head of the Department of Organic Chemistry at Semmelweis University, whose tireless efforts have made this interdisciplinary work possible. I am extremely grateful to Péter Antal, my former master's thesis supervisor and the leader of the Computational Biomedicine research group at Budapest University of Technology and Economics, who gave me indispensable advice from my undergraduate years until the present.

I would like to thank Krisztina Palesits and Marianna Machata for their valuable advice in language issues; Márta Hegyi, Bence Bolgár, Gergely Temesi, András Gézsi, Péter Marx, Péter Sárkozy and the whole Computational Biomedicine Research Group at the Budapest University of Technology and Economics for their scientific support and friendship.

My colleagues in the Institute of Organic Chemistry of the Semmelweis University are sincerely thanked, especially Balázs Balogh for his invaluable help with the Schrodinger Suit software package.

I would like to express my gratitude to my colleagues at the STADIUS Research Group at KU Leuven, especially Jaak Simm, Griet Laenen, Pooya Zakeri, Sara Elshal, Amin Ardeshirdavani, Daniele Parisi and Yves Moreau, my promotor at KU Leuven for that incentive international environment where I worked during the finishing of this thesis; I am also indebted to Prof. Norbert Haider for the possibility to work in his group at the University of Vienna.

I am grateful to my family, my friends, and all who helped me to maintain a peaceful and inspiring atmosphere to work.

Leuven, 2016