Prediction of Transport, Pharmacokinetics, and Effect of Drugs

Inauguraldissertation

zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von

Felix Hammann
aus Deutschland

Basel, 2009

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät
auf Antrag von:

Prof. Dr. Jürgen Drewe
Prof. Dr. Jörg Huwyler

Basel, den 23. Juni 2009

Prof. Dr. Eberhard Parlow
Dekan

“Realism can break a writer’s heart.”

Salman Rushdie, in ‘Shame’

“Elämä on laiffii.”

Matti Nykänen

“Computer says: no.”

David Walliams

Acknowledgments

The three years that went into this thesis were eventful, to say the least, and I can hardly imagine any other place where work as a Ph.D. student would have been more rewarding, productive, and fun. Of course, my supervisor Prof. Dr. Jürgen Drewe deserves the lion's share of credit for this. His scientific creativity, his ability to look beyond biomedicine, and his unique talent to build and foster a focused yet relaxed atmosphere in a crowded lab with people of diverse backgrounds are probably one of a kind.

Dr. Heike Gutmann is the head of lab anyone would wish for. Her unpretentious and warm personality often made me forget she was actually my direct boss. I will miss working with her as much as I will miss the coffee breaks. Uschi Behrens has often been referred to as the 'gem of Lab 411'. Rightly so. She shared her competence, skill, and long experience at the bench patiently and more often than not showed me where the pointed end of the pipette was. Fondly remembered for many of the same reasons are those who shared the bench with me. Dr. Angelika Graber-Maier and Dr. Birk Poller were uncomplicated, supportive, and have a sense of humor which generally made work a blast. I will certainly remember the field trips to the Black Forest and to Paris, be that for the Dalton gang or Louis de Funès. Clinical work was always together with Dr. Oliver Kummer, an apt clinician and colorful personality in and out the workplace, to whom I owe many insights (such as the proper spelling of 旨味). Thanks go to the members of the defense committee. Prof. Dr. Jörg Huwyler was kind enough to participate in the examination and Prof. Dr. Peter Hauser acted as chairman.

During these three years, I had a number of supporters who took me into their departments: Prof. Dr. Dr. Stephan Krähenbühl who heads the Dept. of Clinical Pharmacology and Toxicology. Dr. Christoph Helma, to whom I am grateful not only for the time at the University of Freiburg but also for essential help in software design and understanding the principles of QSAR. And, of very special note, Prof. Dr. Christoph Beglinger, head of the Department of Gastroenterology, who generously enabled me on numerous occasions to continue the work. I thank him as well as his scientific team, Robert Steinert and Anne Christin Gerspach. Klingelbergstrasse was where I finally wrapped things up with Filterkaffee and a dog filled with helium.

Dr. Dr. Petr Hruz must be mentioned for getting me started in the first place. He has been supportive and collegial in a way I could not have expected. In the same vein, I am grateful to Dr. Christian Zimmermann who helped greatly in securing grants. Industrious and diligent people contributed data in the course of their diploma and medical theses. These are Ursi Jecklin, Nadine Vogt, Stefania Guercioni (during the nicotine trial with Prof. Dr. Georgios Imanidis), and Dr. Ulli Baumann. Special mention should also go to the many people around the lab and the various departments: Dr. Michael Bodmer, Dr. Yolanda Brauchli, Dr. Sabin Egger, Evelyne Furger, Dr. Manuel Haschke, Zorana Radic, Evelyne Rudin, Bea Vetterli.

Finally, family and friends were most patient with me: most notably my parents, Barbara and Heinz, who continuously supported me throughout the thesis and the work before. And so did Helen, not only by proofreading and making coffee.

Abbreviations

ABC ..... ATP Binding Cassette
ABD ..... ATP Binding Domain
ADME ..... Absorption, Distribution, Metabolism, and Excretion
AI ..... Artificial Intelligence
ARNT ..... Aryl Hydrocarbon Receptor Nuclear Translocator
ASCII ..... American Standard Code for Information Interchange
ATP ..... Adenosine Triphosphate
ANN ..... Artificial Neural Network
BBB ..... Blood-Brain Barrier
BCRP ..... Breast Cancer Resistance Protein
CAR ..... Constitutive Androstane Receptor
CART ..... Classification And Regression Trees
CCR ..... Corrected Classification Rate
CHAID ..... Chi-square Automatic Interaction Detector
CI ..... Confidence Interval
CNS ..... Central Nervous System
CPSA ..... Charged Polar Surface Area
CV ..... Cross Validation
CYP ..... Cytochrome P
DBS ..... Drug Binding Site
DME ..... Drug Metabolizing Enzyme
DTI ..... Decision Tree Inference
HBA ..... Hydrogen Bond Acceptor
HBD ..... Hydrogen Bond Donor
HTTP ..... Hypertext Transfer Protocol
HTS ..... High-Throughput Screening
IC_{50} ..... Mean Inhibitory Concentration
ID3 ..... Iterative Dichotomizer 3
kNN ..... k-nearest Neighbor
LD_{50} ..... Mean Lethal Dose
LOO ..... Leave-One-Out Cross-validation
MDL ..... Minimum Description Length
MDR ..... Multi-Drug Resistance
ML ..... Machine Learning
MLP ..... Multi-Layer Perceptron
MRP ..... Multidrug Resistance associated Protein
NSAID ..... Non-Steroidal Anti-Inflammatory Drug
OAT ..... Organic Anion Transporter
OR ..... Odds Ratio

Abbreviations

P-gp	P-glycoprotein
PPARα	Peroxisome Proliferator Activated Receptor α
PXR	Pregnane X Receptor
QSAR	Quantitative Structure Activity Relationship
QPSR	Quantitative Structure Property Relationship
RF	Random Forest
ROC	Receiver Operating Characteristic
RXR	Retinoid X Receptor
SABL	Smallest Associated Binary Label
SJW	St. John's Wort
SMARTS	SMILES Arbitrary Target Specification
SMILES	Simple Molecular Input Line Specification
SNP	Single Nucleotide Polymorphism
SQL	Structured Query Language
SSSR	Smallest Subset of Rings
SVM	Support Vector Machines
TPSA	Total Polar Surface Area
UDP	Uridine diphosphate
UDPGlcUA	UDP-glucuronic acid
UGT	UDP glucuronosyltransferases
VDWSA	van der Waals surface area
XML	Extensible Markup Language

Table of Contents

1. Summary

Prediction and modulation of pharmacokinetics and the effects of drugs is a major concern in drug discovery, drug safety, and clinical practice. The projects in this thesis span the entire range of levels upon which these issues can be explored, ranging from *in silico* to *in vivo* evaluations and molecular, cellular, organ, and systemic phenomena. Drug discovery has come a long way, from a mostly serendipitous endeavor to the highly focused process in today's global pharmaceutical companies, where large numbers of candidate substances are routinely weeded out to arrive at a few promising drug leads.

Effective wet lab screening tools (such as high-throughput screening (HTS)) exist to evaluate compounds that exist physically. Increasingly, however, chemical libraries are designed using combinatorial chemistry – by applying a pre-defined set of permissible reactions to a few scaffold structures, by creating databases of compounds with a certain degree of similarity to members of a successful drug class, or any other such method. These virtual libraries require virtual screening tools. The first set of projects in this thesis presents such tools. They assess pharmacokinetic and toxicological profiles using machine learning methods. The second set of projects deals with the clinical implications of individual pharmacogenetic differences and ways of predicting or circumventing them.

The first study outlines the development of decision tree induction (DTI) models to evaluate interactions with P-glycoprotein (P-gp, MDR1, ABCB1). This efflux pump protein is a major factor in the elimination of xenobiotics from cells. Substrate, inhibitor, and inducer properties are predicted by these models with an accuracy of 77.7 % (substrates), 86.9 % (inhibitors), and 90.3 % (inducers). Furthermore, the study shows the superiority of the DTI algorithms CHAID and CART over the more widely used work-horse algorithm C4.5, and the utility of lipophilic distribution coefficients that take into account differing states of ionization depending on the pH of the environment.

In the second study, interactions with the Cytochrome P_{450} (CYP) family of enzymes are predicted. These enzymes are involved in phase I metabolism of most drugs, during which compounds are not only detoxified and prepared for elimination but also active metabolites are formed. Because of the promiscuity and cross-selectivity of the different CYP isoforms, they are heavily involved in complications arising from co-administration of multiple drugs or seemingly innocuous natural substances (e.g. St. John's wort, grapefruit juice). Models were built for substrate, inhibitor, and inducer activities for CYP 1A2, CYP 2D6, and CYP 3A4 using DTI, the k-nearest neighbor algorithm, random forests, artificial neural networks, and support vector machines. Predictive accuracies were very high (81.7 to 91.9 % for CYP 1A2, 89.2 to 92.9 % for CYP 2D6, and 87.4 to 89.9 % for CYP3A4). The commonly held hypothesis that P-gp substrates are CYP 3A4 substrates was evaluated using data from the first study. Both datasets overlap for only 84 compounds and agreement was moderate at 45%.

The third study takes a more global approach and discusses the development of models for broad classes of adverse drug reactions (in the central nervous system, liver, kidney, and allergic potential). Here, the adverse drug reactions of over 500 drugs were determined from the drug register of Switzerland and

classified according to the categories mentioned above. The resulting models perform very well (88.0 to 89.7 % for the CNS, 87.1 to 90.2 % for the liver, 84.7 to 88.6 % for the kidney, and 78.4 to 78.9 % for allergic potential) and can serve as valuable tools not only in early drug discovery but also in pharmacovigilance. A comparison with the data on P-gp obtained in the first study of this thesis indicates that compounds extruded by P-gp do not cause less CNS related adverse drug reactions.

The fourth project deals with the potential consequences of dysfunction in the efflux pumps P-gp and breast cancer resistance protein (BCRP, ABCG2). Several genetic polymorphisms are known for these proteins, some of which alter phenotypes. Because both proteins are expressed at crucial barriers such as the luminal wall of the intestine or the blood brain barrier, decrease in activities is thought to influence disease susceptibility and severity in inflammatory bowel disease (IBD, i.e. Crohn's disease and ulcerative colitis). Furthermore, they influence success of pharmacotherapy because many drugs used in the treatment of IBD are transported by or alter activity of these pumps. Genetic constitution also influences activity and knowledge the individual pharmacogenomic profile could help predict response to pharmacotherapy. In this study, the prevalence of two single nucleotide polymorphisms (SNPs) of P-gp (C3435T, G2677T) and one SNP of BCRP (C421A) were assessed in peripheral blood in healthy volunteers and patients newly diagnosed with IBD, both from the Swiss population. All three SNPs are known to result in a decrease in activity. The rationale behind this study was therefore that such constitutional changes in efflux pump activities may predict disease susceptibility for IBD. While no statistically significant results could be obtained, there are discernible trends towards BCRP 421A ($p < 0.18$), MDR1 2677T ($p < 0.27$), and the wild type allele MDR1 C3435 ($p < 0.46$) in patients with ulcerative colitis. Also in these patients, the haplotypes MDR1 3435CC / BCRP 421CC (X^2: 1.0142, $p < 0.30$) and MDR1 2677G / BCRP 421A (X^2: 1.5615, $p < 0.22$) were more prevalent, although, again, with no statistical significance. IBDs are complex multifactorial diseases and any single SNP is unlikely to serve in reliably predicting susceptibility. However, the study showed the promise of these SNPs and haplotypes derived from them, and a repeat in a larger sample from the Swiss population may reveal stronger associations. Also, because the genotype analysis was performed in peripheral blood (obtained by venipuncture instead of more costly and complex intestinal biopsies), such a study may yield an easily applicable tool for predicting response to pharmacotherapy with P-gp and BCRP substrates.

The fifth study deals with a rare genetic variation in the phase II metabolizing enzyme uridine 5'-diphospho-glucuronosyltransferase 1A1 (UGT 1A1) and the ensuing Crigler-Najjar syndrome type II (CN II) in a Caucasian patient. The patient was suffering from the consequences of hyperbilirubinaemia, which are essentially cosmetic in nature. CN II patients still have residual UGT 1A1 activity (compared to CN I, which, prior to the availability of phototherapy often resulted in neonatal death due to kernicterus). Currently, patients are treated with phenobarbital which binds to the phenobarbital-responsive enhancer module of UGT 1A1. Phenobarbital is a barbiturate with a small therapeutic window and many adverse effects such as drowsiness, impaired motor function, sexual dysfunction, and dependency. Less toxic alternatives are therefore desirable. The residual activity in CN II could be enhanced in a variety of other ways, for example by translational activation via the pregnane X receptor (PXR), the glucocorticoid receptor (GR), or the aryl hydrocarbon receptor (AhR). It is therefore possible to predict that since some

degree of UGT 1A1 activity is present in the patient, translational activation can increase UGT-1A1-mediated metabolism. This study tested this prediction for the PXR activator hyperforin (one of the constituents of St. John's Wort (SJW)). SJW is available in standardized extracts, generally well tolerated (photo-toxicity, fatigue, and gastrointestinal discomfort are the most common adverse reactions), and cost-efficient. The study showed a significant decrease in plasma bilirubin levels in the patient and suggests that SJW is an interesting therapeutic alternative.

The sixth project is a proof-of-principle study showing the feasibility of transdermal pulsatile administration of nicotine in heavy male smokers. Nicotine substitution is an important therapeutic tool in weaning patients from tobacco and it is predicted that mimicking the pharmacokinetic profile of cigarette smoking (short bursts of nicotine exposure resulting in acute peaks in plasma and fast elimination) may give better outcomes (abstinence from tobacco) than conventional modes of administration (e.g. reservoirs in transdermal patches which give long-lasting nicotine plasma levels or nicotine gum which, while giving the patient more control over release than patches, still continuously releases fractions of its content). For the study, volunteers were subjected to three increasing doses of nicotine from a prototype computer controlled device and showed statistically significant peaks in nicotine plasma levels for the highest dose. Adverse events, esp. on the skin, were minimal and receded shortly after the device was removed. The device is a promising way of nicotine substitution and could also be applied to other settings where transdermal patient-controlled delivery is desirable (e.g. opioids in pain control).

In an isolated seventh project, a fully automated system was established for a general unknown screening for toxic substances in serum and urine using liquid chromatography / mass spectrometry technology. A library of over 350 compounds along with spectra is presented as well as a computer program which uses this library to identify substances on-line.

2. Aim of thesis

The main aim of this thesis is the exploration of different techniques to predict or modify transport, pharmacokinetics, and the effect of drugs. Much emphasis lies upon machine learning methods for the study of metabolic pathways, toxicological aspects, and ultimately drug discovery and safety. The role of pharmacogenetics in the pathogenesis of inflammatory bowel disease and Crigler-Najjar syndrome type II and in drug metabolism is also discussed.

The first part of this thesis presents numerical models for the prediction of interactions with P-glycoprotein and the cytochrome P_{450} system of enzymes. Also presented are predictive models for general classes (central nervous toxicity, hepatotoxicity, nephrotoxicity, and allergic potential) of adverse drug reactions. The second part features projects on the pathology of drug metabolism in inflammatory bowel disease (concerning the relevance of single nucleotide polymorphisms and related haplotypes in P-glycoprotein and breast cancer resistance protein) and Crigler-Najjar syndrome type II. Strategies for modulating metabolism in the latter are discussed. An alternative way of transdermal drug delivery is presented in a study of the effect of pulsatile nicotine release by a computer-controlled prototype mounted on the skin of male heavy smokers. In an isolated project, a fully automated system for the determination of toxicologically relevant compounds in liquid chromatography / mass spectrometry is described.

3. Introduction

3.1. Computational intelligence in drug discovery

3.1.1. Approaches to drug discovery

Serendipity (or, less flatteringly put, accident) accounts for many advances and quantum leaps in pharmacology. The discovery of penicillin is attributed to Sir Alexander Fleming's accidentally contaminated bacterial cultures [1], oral contraceptives were the accidental by-product of Carl Djerassi's synthesis of progesterone [2] and the accidental discovery of the psychedelic effects of lysergic acid diethylamide (LSD) during a bicycle ride home [3] was commemorated by its inventor Albert Hofmann until his death with the celebration of the annual 'Bicycle Day'.

A more directed approach, that of bounded rationality, lies in the tapping of the wealth of pharmacological experience in traditional medicine and herbal remedies (the initial discovery of which is also attributable to serendipity) by isolating active ingredients. Willow tree bark, for example, has long been appreciated for its antipyretic and analgetic effects, first leading to medicines based on a bark extract in the 19th century, and culminating in the preparation of acetylsalicylic acid in the 1890s and subsequent marketing of the drug under the name 'Aspirin' by Bayer [4].

As understanding of the molecular pathology of disease and toxicology increases, so does the possibility to create models of specific sites of action and how ligands (i.e. drug candidates) interact with them. For these models to work, a lot of information on a lot of different structures and mechanisms has to come together, and although there is an ever growing body of structural and mechanistical information, there are still only a few cases where this paradigm has led to tangible results. [1]

In situations where targets of interest (such as receptors or transporters) have been identified, one could try and identify drug candidates with high-throughput screenings (HTS), i.e. by confronting them with a large set (10^4 to 10^5) of available compounds in an in-vitro assay. Continuous minimization of assay technology, e.g. the move from 96- over 384- to 3456-well micro-titer plates, reduction in variance of proven strategies such as radio-ligand binding, and automation through use of robotics allow ever higher numbers of substances to be screened within a day's time.

However, with a chemical universe (the set of potentially interesting compounds) estimated to number 10^{62} or even higher [5], it remains questionable whether a brute-force strategy such as HTS is efficient to find the metaphorical needle in the haystack. Nevertheless, one can enhance the output of HTS (and even replace it in early stages) by testing only 'promising' subsets of compounds.

[1] The anti-glaucoma agent dorzolamide (1995) and the tyrosine kinase inhibitor imatinib (2001) are examples of successful rational drug design.

This can be achieved through studies of quantitative structure activity relationships (QSAR) where a certain biological property is measured for a range of different structures in the hopes of establishing a relationship between structure and activity. Insights into structural requirements are then applied to a wider range of compounds of unknown activity, even compounds that only exist in a virtual chemical library whose entries have not been synthesized yet. In a way, this testing of more or less random compounds can be seen as taking a full circle back to the early days of drug discovery.

3.1.2. Quantatitive Structure Activity Relationships (QSAR)

The fundamental concept in Quantatitive Structure Activity Relationships (QSAR) studies is the Similar Property Principle, i.e. the assumption that similar structures have similar properties or activities [6]. The goal is to predict compound activities by analyzing their structures (*in silico*) instead of making explicit measurements *in vitro* or *in vivo*. Correlating pharmacodynamical or toxicological effects solely with structure while disregarding the molecular mechanisms at the target level may seem like an oversimplification at first. However, if a complex in question (e.g. a receptor or enzyme) is able to distinguish between active and inactive molecules, structure-based models should in theory be capable of doing the same [5].

3.1.2.1. Historical background

Pioneering work by Hermann Kopp dates back to the mid-19th century [7-9], when he described changes in boiling points and atomic volumes in compounds that were largely part of a homologous series of alkanes. In fact, this may be seen as the first appearance of the additivity principle [10], stating that each of several structural features of a molecule makes a separate contribution to a given property. Kopp formulated these relationships in algebraic equations and matrices using physicochemical properties.

Credited with first proposing a mathematical relationship between structures and their activity are Crum-Brown and Fraser in a series of papers published in 1868 and 1869 [11]. Studying the curare-like properties of a series of quaternized strychnines, they came to note the importance of the quaternizing group. In their words, physiological function Φ (or activity) follows from a compounds constitution C (or structure):

$$\Phi = f(C)$$

Changing the constitution (ΔC) will bring about a change in function ($\Delta\phi$). At the time, however, it was thought that chemical structures could not be subjected to mathematical analysis. This perception changed with a landmark of QSAR studies, the lipoid theory of anesthesia, put forward by Overton [12] (building on work by Meyer [13]), in which he showed that the potency of aqueous anesthetics in tadpoles correlates with their partitioning in water and lipophilic phases [2], P':

[2] Overton used olive oil. Today, lipophilic partitioning is usually assessed with an octanol phase (see below and in section 4.1.7.4).

$$P' = (\text{concentration in organic phase}) / (\text{concentration in water})$$

It was later shown [14] that the effective anesthetic concentration *C'* in tadpoles is fairly constant:

$$C' \cdot P' = const. \approx 0.05$$

Expressed in its logarithmic form, the equation is known as the Overton-Meyer relationship, the first QSAR ever reported:

$$\log\left(\frac{1}{C'}\right) = \log(P') + const.$$

Apart from formalizing empirical findings, this equation has other important applications. Firstly, one may deduce a structural requirement for anesthetic agents: high lipophilicity. [3] Secondly, it can be used to screen unknown compounds for their narcotic potential by determining their lipid solubility. [4]

3.1.2.2. Origins of modern QSAR: sigma and pi effects

A further milestone occurred in 1937, when Hammett published the equation that bears his name [17]. Hammett correlated substituents of organic acids with their reactivity. In aromatic systems such as the benzene ring, electronic charges are distributed equally. Substituents introduced to the system can delocalize these charges.

[3] It has been shown that this parameter cannot be extended *ad libidum*, i.e. analgetic potency cannot be increased infinitely to achieve ever higher effects. This is nicely summarized in the phrase 'methyl, ethyl, butyl, futile' [15].

[4] Meyer and Overton deduced from this that anesthetic effects occur when an agent reaches a critical concentration in a lipid phase of the body. This theory, however, fails to explain receptor-based analgesia (NMDA, opioid, and $GABA_A$ receptors) or effects on Na^+ and K^+ channels. While lipophilicity today is recognized as an important molecular feature, it is seen as insufficient to provide the global theory of anesthesia it was hoped to [16].

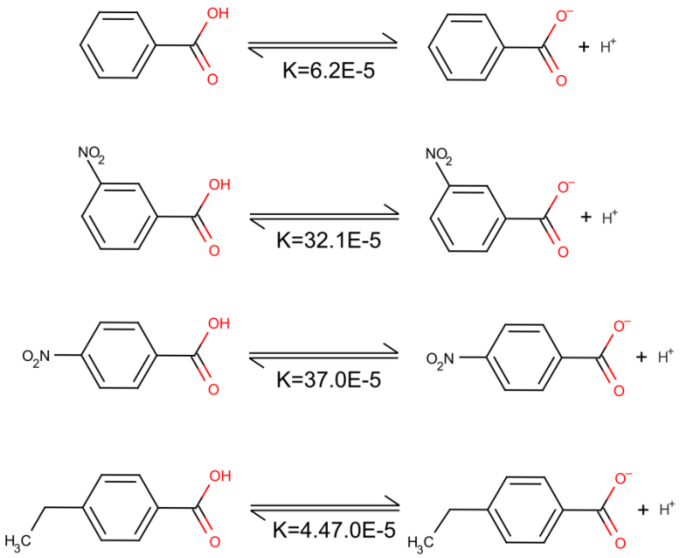


Figure 1 – Changes in the dissociation constant *K* in a series of organic compounds. *K* is influenced by the type and position of substituents.

Consider the series presented in Figure 1, where the first reaction shows the dissociation of benzoic acid with a dissociation constant K of $6.27 \cdot 10^{-5}$. A nitro group introduced in the *meta* position increases the degree of dissociation (by withdrawing charge from the carboxyl group and stabilizing the negative charge in the product molecule), an effect even more pronounced when the group is moved to the *para* position. An ethyl group in the same position, however, decreases dissociation through its electron-donating quality.

Hammett came to realize that similar modifications of other organic compounds produce similar effects. Let the dissociation constant of two distinct compounds be K_0 and K_0', respectively, and K and K' be the respective values for the same compounds substituted with the same functional group in the same position, then the following equation can be written:

$$\log \frac{K}{K_0} = \rho \log \frac{K'}{K_0'}$$

The reaction constant ρ depends solely on the reaction itself (in this case: acid-base dissociation). As the effects of a substituent in this context only depend on the functional group and not on the underlying molecule, the equation is usually written with a substituent constant σ (stating the relative strength) as:

$$\log \frac{K}{K_0} = \rho\sigma$$

With this, Hammett had introduced a concept known as the 'sigma-effect' to physical organic chemistry. His results were quickly applied to, amongst other things, ionization kinetics and biological equilibria.

While working on plant growth regulators in the 1960s, Hansch and Muir sought to model auxin effects using Hammett-type equations [18], assuming that electron density at the *ortho* position in phenoxyacetic acids is responsible for their effects in plants. Surprised by the low predictive accuracy of their results, they surmised that, while their models were correct, the substances were not getting to the target in the first place. As concentrations within the plant cell directly depend on substances' ability to cross several layers and membranes, Hansch and Muir looked for a suitable measure of permeability.

They found it in a compound's partition coefficient π between 'water and a relatively non-polar solvent' and chose 1-octanol as the lipid phase. [5] The equation for the 'Hansch approach' shows parallels to Hammett's equation:

$$\log\frac{1}{C} = a\pi + b\pi^2 + \rho\sigma + dE_s + const.$$

Here, *C* is the concentration required to produce the desired effect, the coefficients *a*, *b*, *c*, and *d* are determined with regression analysis for each study, σ and ρ characterize substituent and reaction-specific characteristics, and E_s is a modification [6] of Hammett's σ. Hansch and Muir's contribution was two-fold. Firstly, they introduced a hydrophobic substituent constant, π, to account for rate-determining effects by distribution in biological systems. Secondly, they established partitioning in 1-octanol and water as a parameter. Their approach therefore uses hydrophobic, electronic, and steric attributes of molecules to predict effects. This is seen by many as the birth of modern QSAR studies [10, 20, 21]. [7]

A different route was proposed in 1964, when Free and Wilson provided a mathematical analysis of additivity schemes [22]. For the sake of their method, they assumed [8] that the biological effects in a series of homologous compounds are comprised of

1.) a constant contribution of the scaffold structure, and

[5] This was done, amongst other reasons, because 1-octanol has a relatively long aliphatic (non-polar) chain and hydrophilic head, much like the lipid monomers that make up biological membranes.

[6] The modification was proposed by Taft in 1956 [19] and is based on the rate of hydrolysis in esters of the type $X\text{-}CH_2COOR$.

[7] To be precise, the efforts before this were *QSPR* studies, i.e. correlations were established between structures and *properties*. Hansch and Muir's contribution was recognizing the importance of absorption to achieve biological effects. QSAR is currently the most widely used term for both.

[8] While the first assumptions are more or less easily accepted, the third assumption is clearly disproven by, amongst other things, Hammett's sigma effect.

2.) a constant contribution (positive or negative) of substituents, and
3.) that no interactions exist between substituents themselves or the scaffold.

Free and Wilson took an observable biological response variable (LD_{50} in analgesic compounds) and noted the changes when substituents are added to a given atom. They then compared the change in the response variable to its average value and were thus able to numerically estimate substituent effects without the need for time-expensive physicochemical calculations. Unfortunately, models built with this method fail to predict activities for substituents and configurations not included in the original data. [9]

3.1.2.3. QSAR today

Crum-Brown and Fraser [11] stated the following requirements for the creation of QSAR models: a reference set of compounds is needed, based on a selection criterion (an activity), and complemented with values describing the compounds so they can be analyzed mathematically. This still holds true today [23].

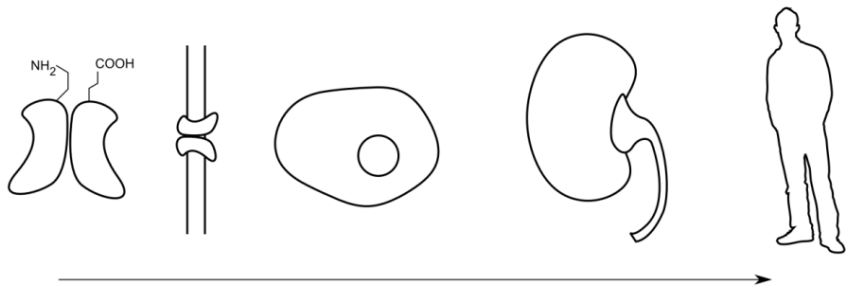


Figure 2 – Possible target activities in QSAR studies range from molecular and receptor level kinetics to cellular toxicity and organ effects up to actions observable in the organism as a whole (adapted from [20]).

The promise of predicting activities before spending time and money on expensive synthesis and assays has seen QSAR methodology being applied to a wide array of targets. Especially intriguing is the fact that one need not have an understanding of how compounds achieve their effect in order to create successful models (although it helps). All that is needed is an observable quality (receptor activity, cellular toxicity, systolic blood pressure, to name but a few examples) and a set of compounds to relate it to (Figure 2).

Once a data set of compounds and activities has been collected, it is extended with values describing the compounds (so-called ‘descriptors’). The type of descriptors used can serve to roughly categorize the studies themselves, i.e. their dimensionality (Table 1). The historical approaches described in the preceding sections are 1D or 2D, and purely mathematical models will usually not go beyond the third dimension. An example of higher-dimensional studies is the generation of three-dimensional computer

[9] Or: the training set (as will be seen below).

models of pharmacophores and the computer-aided fitting of ligands to binding sites (e.g. [24]). The need for the interaction by a human expert makes these approaches unsuitable for high-throughput screenings of virtual libraries. They do, however, provide mechanistical interpretations and models that statistical analysis will not always produce.

Dimensionality	Correlations
1D	Physicochemical properties (weight, lipophilicity, etc.)
2D	Structural motifs, functional groups
3D	Actual three-dimensional structure
4D	As 3D, but including conformational changes
5D and beyond	As 4D, but including induced fit models, solvation models, etc.

Table 1 – Definition of QSAR dimensionalities and what they correlate with (adapted from [25])

After descriptors are available, a variety of methods can be used to create a model. The procedure generally follows the scheme given in Figure 3: the available data is split into a training set to create the model and test set to estimate performance on unseen compounds. If the predictive accuracy meets expectations, the model can be applied to compounds with unknown activity.

Before performing a virtual screening on a large database, it may be wise to filter it for cost-efficient compounds and other factors such as drug-likeness. For the latter, rules of thumb have been proposed. The best known might be the 'rule of five', put forth by Lipinski [10] et al. in an analysis [26] of structurally diverse drugs, where they found that a great many of well-absorbed or well-permeating drugs have ≤ 10 hydrogen bond acceptors, ≤ 5 hydrogen bond donors, a molecular weight ≤ 500, and a logP ≤ 5. Filtering based on functional groups may also be applied. [11] A similar set of rules was put forth by Veber et al. [27] in 2002 for oral bioavailability in rats, coming to similar conclusions as Lipinski et al. According to this work, determinants for good bioavailability are ≤ 10 rotatable bonds and 140 $Å^2$ polar surface area (see 4.1.7.4) or ≤ 12 hydrogen bonds.

[10] Lipinski, notably, is an employee of a large pharmaceutical firm, and therefore prone to streamlining of drug discovery processes.

[11] Substances that fail the 'rule of five' (mostly antibiotics, antifungals, and vitamins) are thought by the authors to achieve drug-likeness by their special functional groups. Amending screening parameters can therefore lead to refined results.

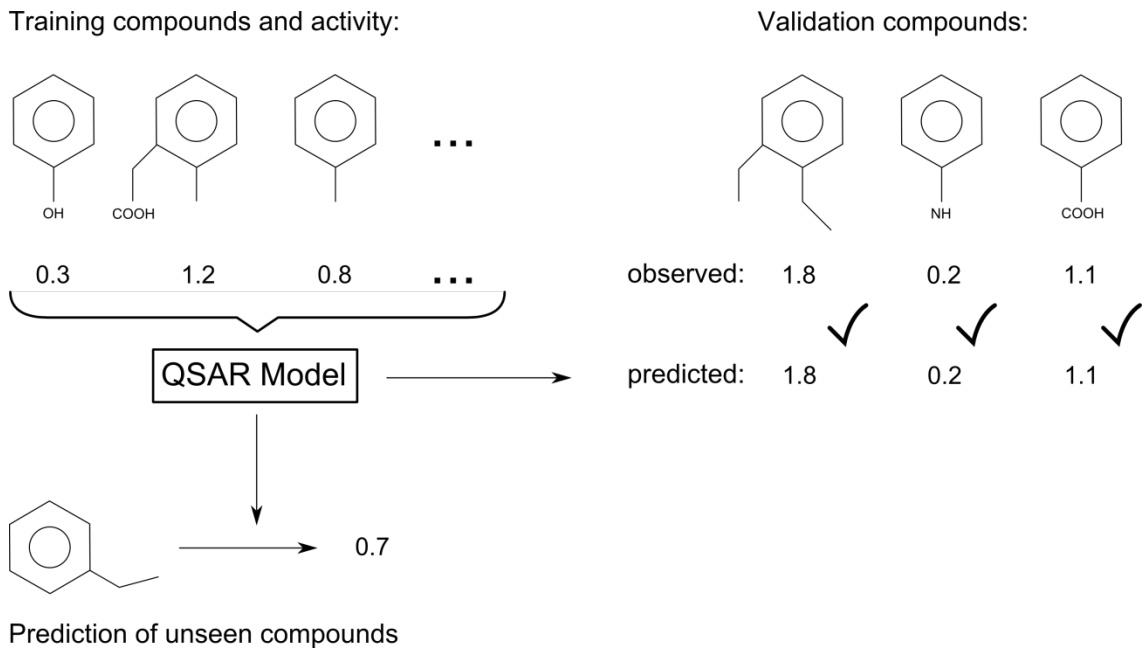


Figure 3 – Development of QSAR models (example). A set of compounds with known activity of interest is split into a training and a validation / test set. The training set is analyzed using a statistical or learning method to produce a model which is applied to the validation compounds. If the model is satisfactory, it can be used to predict activity of unseen compounds.

Compounds selected from a chemical database using QSAR models are called 'hits'. These may be examined further in in-vitro assays to produce 'validated hits' or complemented with similarity searches (see below) of the database. Eventually, what is left are lead candidates. Such compounds receive further expert evaluation and may proceed into clinical trials. [28]

3.1.3. Computational Intelligence

The field of computational intelligence is hard to pin down, with definitions ever changing until first efforts were made to define it at the Dartmouth conference in 1956 as 'artificial intelligence' (AI). [12] A discipline of computer science, it is now widely regarded as the study and design of intelligent agents [29]. Its major goal, the reproduction of general intelligence (or 'strong AI') in machines, seems to still lie far in the future, and its feasibility is not only debated in the scientific community but also philosophically. Always looming over the field is the infamous test formulated by British mathematician Alan Turing in 1950 in the philosophical journal 'Mind' [30]. Turing stated that a machine should be considered intelligent if a human volunteer in a conversation over a typewriter terminal is unable to tell whether the conversation is with another human being or with a machine. To date, no system is considered to have passed this test.

[12] The term was coined by John McCarthy when he proposed that intelligence can be described so accurately that a machine can reproduce it.

Research in the quest for this goal, however, has led to a wide variety of tools that are mostly statistical in nature (an overview is given in Section 4.2). Pattern recognition is an especially intriguing subfield of AI when it comes to drug discovery. The discipline deals with the classification of objects based on *a priori* knowledge or statistical evaluations. Classifications can then be passed to pattern matching systems and be applied to unseen objects [31]. Obvious applications are computer vision (i.e. the automated interpretation of scenes and visuals) and optical character recognition to extract text from printed materials.

In drug discovery, pattern recognition can uncover patterns in chemical libraries and associate them to chemical or biological activities. Pattern matching based on classification systems is easily applied to chemical databases in high-throughput screenings (HTS, see below).

3.2. Drug Metabolism

Drug metabolism is the set of changes xenobiotics can undergo after they have been absorbed. Usually, lipophilic compounds are converted to hydrophilic compounds which are more readily excreted by the kidney and less easily re-absorbed in the renal tubuli. Metabolism commonly limits or terminates drug action, although in some instances, metabolites are the active principle. This is desired when, for example, the bioavailability of the parent compound is superior or its toxicity is lower (and the metabolite forms only at the site of action). Many pro-drugs have been discovered serendipitously (e.g. molsidomine, which is metabolized to the unstable active SIN-, which, in turn, releases NO upon decay, or heroin, which is de-acetylated to form active morphine derivatives) [32].

3.2.1. Functions of metabolism

Conventionally, drug metabolism is divided into two phases, which do not necessarily occur in sequence. Chemical alteration such as hydroxylation, oxidation, or reduction are part of phase I metabolism, for example procaine, which is hydrolyzed by pseudocholinesterase to para-aminobenzoic acid (PABA) and diethylamino-ethanol (Figure 4a). [13] An overview of phase I enzymes is given in Table 2. A great number of phase I reactions are carried out by Cytochrome P_{450} (CYP) mono-oxygenase superfamily of enzymes. In humans, these are located on the endoplasmic reticulum or on the inner membranes of mitochondria, and are subdivided into different families. The most important ones in drug therapy are CYP1A2, CYP2C9, CYP2C19, CYP2D6, and CYP3A [32].

In general, phase I reactions introduce (e.g. R-H → R-OH) or unmask (e.g. $R\text{-}CO_2CH_3$ → R-COOH + CH_3OH) polar groups, or convert existing groups (e.g. R=O → RH-OH). These transformations increase solubility in water or prepare compounds for phase II of metabolism.

[13] Deficiency in this enzyme is a recognized disease that usually only becomes apparent when undergoing anesthesia.

Reaction	Enzymes (examples)
Oxidation	CYP-P_{450}, monoamine oxidase (MAO), alcohol dehydrogenase, peroxidases, flavin-containing mono-oxygenase (FMO)
Reduction	NADPH-cytochrome P_{450} reductase
Hydrolysis	Esterases (acetylcholine esterase, pseudocholinesterase)

Table 2 – Examples of enzymes involved in phase I metabolism

Phase II reactions are conjugations that further increase hydrophilicity, typically with conjugates such as sulfate, methyl groups, acetate groups, glutathione, amino acids, or glucuronic acid. Conjugations are performed on parent compounds or their phase I metabolites, i.e. the two phases are complimentary and not mutually exclusive. An example showcasing the different fates of a single molecule is given for acetaminophen (Figure 4b).

A final barrier before compounds reach their site of action are transport systems such as multidrug resistance proteins (MRPs), organic anion transporters (OATs), and P-glycoprotein (P-gp), all of which have been described above. They constitute phase III of metabolism [33].

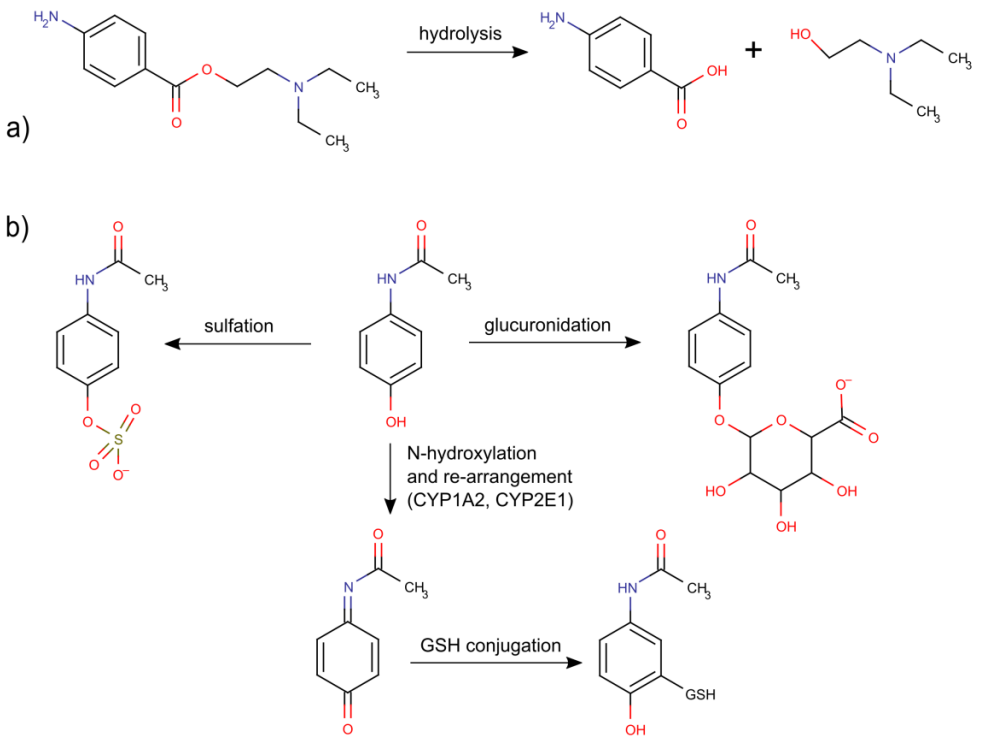


Figure 4 – Metabolism of procaine (a) and acetaminophen (b). Procaine is subject to hydrolysis by pseudocholinesterase whereas acetaminophen is either sulfatized (20-40%), glucuronidized (40%),

or metabolized by CYP enzymes to the toxic N-acetyl-p-benzo-quinone imine (NAPQi) which, in turn, is deactivated by a phase II reaction (conjugation with glutathione-SH (GSH)).

3.2.2. Pathology of Drug Metabolism

3.2.2.1. Genetic constitution

Genetic variations of drug metabolizing enzymes (DMEs) are a source for many pharmacological interactions and adverse events. A metabolizing enzyme family that is receiving increasing attention is that of UDP glucuronosyltransferases (UGT). They act on small lipophilic compounds (e.g. bile acids, steroids, bilirubin, and hormones) and conjugate them with UDP-glucuronic acid (UDPGlcUA). UGTs have been identified in virtually all vertebrates and show a broad specificity [34]. Enzyme deficiency has been well documented for the subtype UGT1A1, which has been implicated in hereditary disorders with decreased bilirubin conjugation (e.g. Gilbert's syndrome [35], and Crigler-Najjar syndrome [36]). Even though many genetic polymorphisms have been identified, their effect on pharmacokinetics remains poorly understood [37].

Of even greater relevance is the CYP superfamily. CYP2C9, for example, metabolizes almost 20% of drugs in clinical use [38], CYP2D6 metabolizes about 15% [39], while CYP3A4/5/7 accounts for a staggering 45-60 % [40]. CYP4 to CYP51 are involved in endogenous pathways and do not play a role in xenobiotics metabolism [39]. Much effort has focused on single nucleotide polymorphisms (SNPs), DNA sequence variations of a single base with a frequency of > 1% in a given population. Most of the studied SNPs are biallelic, although it is apparent that up to three mutant alleles may exist. SNPs have been extensively studied in CYPs and are of importance to the concept of personalized medicine, where not only macroscopic features such as size, weight, or sex are considered, but also genetic constitution. Clinically relevant (functional) SNPs have been reported mostly for CYP2D6 and CYP2C9 whereas the broadly specific CYP3A is largely unaffected by them [38, 41].

3.2.2.2. Interactions

PHASE I

CYP activity is not only influenced by genetic constitution but also by dietary and hormonal factors and exposure to xenobiotics and drugs as has been recognized long ago, for example, in phenobarbital, pesticides, and carcinogenic polycyclic aromatic hydrocarbons [42]. Tobacco (CYP 1A2, CYP 2B1) and alcohol abuse (CYP2E1) [43] are common inducers of CYP activity and frequently interfere with pharmacotherapy and clinical trials.

Most CYPs are induced via receptor-mediated mechanisms, commonly by aryl hydrocarbon receptor (AhR), the constitutive androstane receptor (CAR), the pregnane X receptor (PXR), and the peroxisome proliferator-activated receptor α (PPARα). AhR binds to inducing agents and then forms heterodimers with the aryl hydrocarbon receptor nuclear translocator (Arnt) which induce DNA expression. PXR, CAR, and

PPARα also bind inducers but then form heterodimers with the retinoid X receptor (RXR) to induce expression [44].

Inhibition of CYP activity is conferred by an equally broad spectrum of factors and inhibitors are often part of a healthy diet. A well explored example is grapefruit juice, whose ingredients include psoralens, a known inhibitor of CYP 3A4 [45]. While inhibition may have negative effects on drug therapy (by decreasing plasma concentrations of active metabolites or elimination of parent drugs, resulting in higher AUCs and increased toxicity), it also has benefits: less oxidation implies less toxic metabolites and free radicals. [14]

Clear lines cannot be drawn between substrates, inhibitors, and inducers for CYPs, e.g. substrates of a specific enzyme can induce their own metabolism (auto-induction). In principle, three mechanisms of CYP activity modification are recognized: reversible inhibition, mechanism-based inhibition, and induction on the level of gene expression [47]. Reversible inhibition is analogous to competitive agonism in general enzyme kinetics and arises from the competition of two compounds for the active site. More severe are mechanism-based inhibitions, where an intermediate metabolite binds irreversibly to the CYP enzyme, inactivates it, and only de novo protein synthesis can restore earlier levels of activity. This is influenced by many factors, including the availability of alternative pathways and the presence of other substrates, both of which potentially decrease inhibition (allosterically).

As a consequence of all of this, CYP interaction potential is not easily predicted and even proper pre-clinical ADMET assessment cannot rule out all toxicities. [15] In fact, in recent years there have been high-profile withdrawals from market stemming from (idiosyncratic) CYP interactions, as was the case with the antihypertensive agent mibefradril in 1988 [48]. These interactions have therefore become a major concern in clinical therapy and for pharmaceutical companies who routinely assess the interaction potential of their products [49].

PHASE II

Interactions in phase II enzymes are not as well understood. This is especially unfortunate in the case of the UGT superfamily of enzymes, as they are thought to be involved in the metabolism of 35 % of drugs undergoing phase II reactions [50]. Immunosuppressants (e.g. tacrolimus and ciclosporine A), non-steroidal anti-inflammatory drugs (NSAIDs), benzodiazepines, and tricyclic antidepressants appear to be strong inhibitors. Induction of UGT has been demonstrated for several of isoforms and inducing agents include rifampin, phenobarbital, and phenytoin [51]. It should be noted that there is strong cross-selectivity

[14] Somewhat ironically, the positive antioxidant effects of catechins such as epigallocatechin gallate in green tea [46] are sometimes counteracted by the presence of pesticides on green tea leaves which may induce CYP activity and the generation of radicals.

[15] Several cases are known in which *in vitro* assays show CYP 3A4 inhibition but *in vivo* data show induction. Imidazoles (e.g. clotrimazole) and chlorpromazine are two examples. [47]

with members of the CYP superfamily. This seems reasonable as both systems often act in concert and is supported by recent findings that show homology in transcription factors regulating both families [52]. Cross-selectivity within UGT enzymes and other families makes it difficult to assign specific isoforms to a given agent. Such redundancy allows for functional compensation when one system fails.

3.3. Drug Transport and Metabolism in the Central Nervous System

3.3.1. Blood-brain barrier (BBB)

The blood-brain barrier (BBB) is a functional and physical barrier between peripheral circulation and brain tissue. Its co-evolution with the central nervous system (CNS) is seen not only in humans but also in other vertebrates and even insects [53, 54]. The concept of the BBB was developed well over a century ago, starting with Paul Ehrlich's [16] observation in 1885 that aqueous dyes introduced into peripheral circulation do not stain the brain [55]. Edwin Goldmann, one of his students, repeated the experiments, but injected dye into the subarachnoid compartment of the brain, thereby staining only the brain [56]. The special role of the vascular endothelium was first recognized in 1934 [57] and three decades later, electron microscopy revealed tight junctions as the morphological correlate of the barrier function [58, 59]. The necessity of active transport was proposed in the 1940s [60] on the basis that, given the existence of a virtually impermeable barrier, special transport mechanisms must also exist to supply the brain with nutrients. Today, several membrane transporters are recognized (see below). In fact, catalyzed transport in the form of carriers and transporters and lipid-mediated transport of small lipophilic molecules are considered the only means of effectively passing the BBB [61].

3.3.1.1. Structure of the blood-brain barrier

The overall surface spanned by the blood-brain barrier (20 m^2) is enormous [62]. Cerebral capillaries have the smallest caliber in the entire vascular system (3-7 µm) and are densely packed, with inter-capillary distance as low as 40 µm [63]. The endothelium of capillaries is a tightly woven single-layer stratum resting upon a basal lamina with tight junctions at the lateral interfaces, thereby limiting paracellular passage (Figure 5) [64].

Pericytes, undifferentiated mesenchymal cells capable of transforming into fibroblasts and smooth muscle as well as performing macrophage-like functions, line the capillaries, providing physical support and aiding in angiogenesis [66, 67]. Astrocytic glia cells also communicate with the capillary system. They give metabolic support to neurons and the endothelium, are involved in nervous system repair, and modulate neural transmitter uptake via their connections to synapses. Together with synaptic nerve endings, they make up the structural component of the BBB [68].

[16] This was not Ehrlich's only contribution to medicine. He is credited with first describing mast cells and erythroblasts, formulating the side-chain theory in immunology, and founding the era of chemotherapeutics with his development of arsphenamine (Salvarsan), the first pharmacological therapy of syphilis.

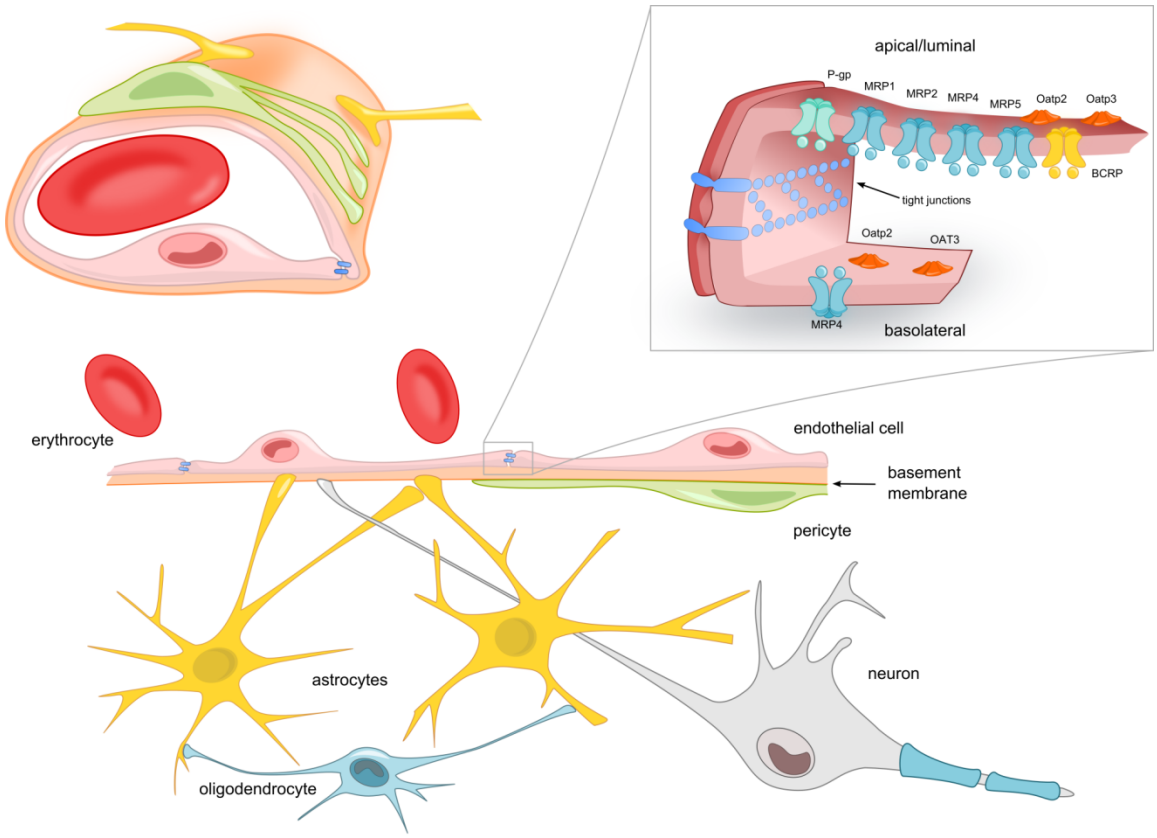


Figure 5 – Structure of the blood brain barrier: cross section of a capillary, the surrounding tissue, and the most important transporters: P-gp (ABCB1), BCRP (ABCG2), multidrug resistance-associated proteins (MRP1, MRP2, MRP4, MRP5), and several organic anion transporters (Oatp2, Oatp3, Oat3) [65].

3.3.1.2. Drug transport over the BBB

Drug transport over the BBB can be grouped into three distinct mechanisms: carrier-mediated (vitamins, thyroid hormones), receptor-mediated (larger molecules such as insulin), and efflux transport (removal of xenobiotics and toxic compounds) [69]. Until today, the mRNA of about 15 transporters of several families has been located at the BBB or detected in BBB cell lines [70, 71]. Examples include P-glycoprotein (P-gp, ABCB1), breast cancer resistance protein (BCRP, ABCG2), multidrug resistance associated proteins (MRP1, MRP2, MRP4, MRP5), and several organic anion transporters (Oatp2, Oatp3, Oat3). Pardridge [61] estimates that these represent about half of the actual inventory.

Efflux transporters are mostly located on apicoluminal surfaces. Some, e.g. Oatp2 or MRP4, are also found basolaterally. This is not well understood, leading to the hypothesis that basolateral efflux pumps pair their function over a gradient with apicoluminal counterparts to move anionic xenobiotics from the CNS to the bloodstream [72].

3.3.1.3. P-glycoprotein (P-gp)

P-glycoprotein (P-gp) is a well known mammalian efflux transporter that is expressed in many human organs with secretory or barrier function, including the liver, the placenta, and the brain [73]. It is encoded for by the MDR gene family, and two isoforms are known in humans: MDR1 (ABCB1, or P-gp) and MDR2 [74], which is of importance for biliar phospholipid secretion in hepatocytes [75]. Other isoforms are found in rodents [76].

MDR1 is located on chromosome 7q21.1, and spans over 100 kb. However, MDR1 mRNA has a size of 4.7 kDa, implying that only a small percentage of the gene actually codes [77]. The product of MDR1 is a 170 kDa transmembrane protein consisting of two homologous parts, both of which are followed by an ATP-binding domain (ABD) that resides in the cytoplasm (Figure 6) [78]. The symmetry of the secondary structure suggests evolution from a tandem repeat of a single gene [79]. Reconstructions based on homology studies with ABC-transporters in *E. coli* and X-ray structures of ABD proteins suggest a tertiary structure with radial symmetry [80].

P-gp is currently the best studied ABC transporter, and to date 50 single nucleotide polymorphisms (SNPs) have been identified, more than half of which reside in the coding region [81-83]. Most SNPs in the coding region of P-gp have a relatively low frequency (<8%) [83]. With a reported prevalence of 34.3 % in Caucasians and 16.9 % in Asians, the SNP C3435T (exon 26) is an exception [84], whose effect on P-gp activity has been shown *in vitro* [85] and *in vivo* [83]. Why this synonymous SNP affects P-gp activity remains unclear. It has been proposed that it is in linkage disequilibrium with non-synonymous SNPs in regulatory regions [83].

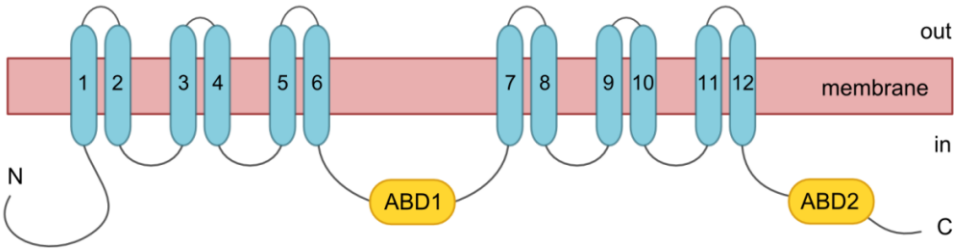


Figure 6 – The 6+6 helical secondary structure of P-gp as proposed by Jones [78]. The ATP-binding domains are labeled ABD1 and ABD2.

The organ distribution and cellular location (apical on endothelia and luminal in the intestine) suggest that P-gp has detoxification function (xenobiotics and toxic compounds). It accepts a wide variety of substrates, ranging from small organic ions to amino acids and even macromolecules like polysaccharides [86, 87]. Other physiological functions are still subject of debate, but are thought to include regulation of apoptosis, stem-cell differentiation, cytokine modulation, and translocation of platelet-activating factor [88].

CLINICAL RELEVANCE OF P-GP

The contribution of P-gp to mucosal barrier function has led to it being implicated in the pathophysiology of several conditions such as Crohn's disease and ulcerative colitis [89, 90]. Studies on polymorphisms of the MDR1 gene, which encodes P-gp, however, have produced conflicting results, and underline both the multifactorial nature of these diseases as well as the relatively low importance of genetic predisposition [77, 91-93]. What can be said is that expression is altered in inflamed gut tissue [92, 94] and that this has therapeutic consequences [92, 95, 96].

MODULATION OF P-GP ACTIVITY

In recent years, attempts have been made at reducing P-gp activity with the goal of developing 'chemosensitizers' that increase therapeutic benefit. Inhibition of P-gp is desirable to increase bioavailability, e.g. of anti-cancer agents, but carries the danger of corrupting the functional BBB by allowing otherwise innocuous compounds to penetrate the CNS and increasing dose-dependent adverse effects. [17] Three modes of interference have been proposed: direct interaction with P-gp, disturbance of ATP-binding and hydrolysis, and disturbance of lipid bilayers, which is thought to influence membrane environment and drug-ligand interaction [98].

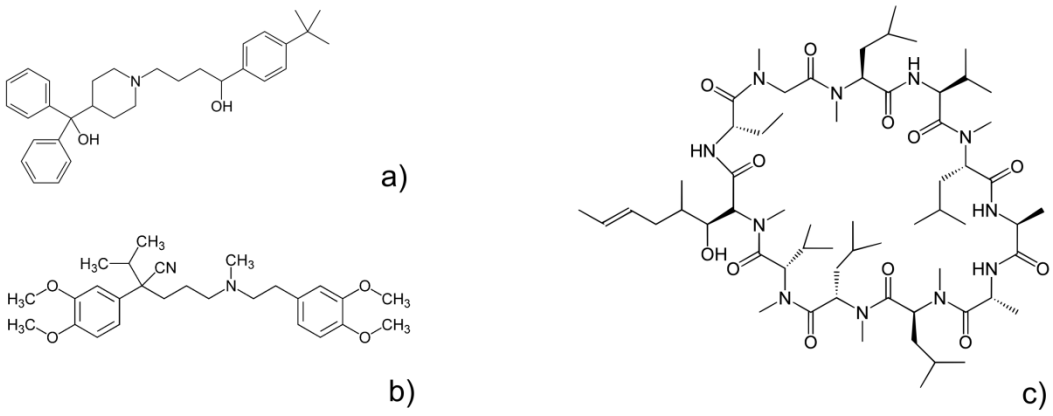


Figure 7 – Structural diversity of P-gp inhibiting compounds: terfenadine (a), verapamil (b), and ciclosporine A (c).

Tsuruo et al. observed that verapamil administration sensitized vincristine-resistant leukemia cells to vincristine and vinblastine [99]. It was quickly shown that not only verapamil [100] but also other structurally diverse compounds (Figure 7) such as ciclosporine A [101] and terfenadine [102] reverse P-gp mediated multidrug resistance. This 1st generation of inhibitors is made up of previously available compounds re-evaluated for their inhibitory effects, and derivatives based on them. These attempts,

[17] Conversely, this may be desirable in order to increase drug concentrations in the CNS, e.g. in anti-epileptic therapy [97].

however, were thwarted mostly by unacceptable toxicity profiles (such as dose-limiting cardiotoxicity for verapamil).

High-throughput screenings formed the basis for 2nd generation inhibitors such as PSC-833 and VX-710 [103, 104]. Peripheral toxicity was lower than in the 1st generation, as was the ensuing dosage of anti-cancer drugs necessary to reach therapeutic plasma levels. However, so was overall efficacy. This apparent paradox was shown to be due to an overlap of CYP3A and P-gp substrate affinity [105] and concomitant regulation of the two metabolizing systems [106].

The current 3rd generation is the product of combinatorial chemistry and characterized by their low mean inhibitory concentration (IC_{50}) [80]. GF120918 (elacridar) [107], LY335979 (zosuquidar) [108], XR9576 (tariquidar) [109], and OC144-093 (ontogen) [110] are notable examples. Although the pharmacophoric characteristics of P-gp are not well understood, basic rules of thumb exist. Wang et al., for example, reported in a QSAR analysis that effective inhibitors should have a logP value of ≥ 2.92, carry at least one tertiary basic nitrogen, have a molecular axis of ≥ 18 atoms, and a high energy of the highest occupied orbital (E_{homo}) [111].

In general, all of these inhibitors have not held their promise in clinical trials. This is probably because of drug resistance inferred by other mechanisms [112] (other efflux transporters, enzymatic degradation, alternate metabolic pathways in cancer cells, decreased permeability, or insufficient concentrations at the target site, to name but a few). The search for specific P-gp inhibitors remains an attractive goal, although specificity might better be extended to other drug metabolizing enzymes so as to cover alternative pathways of degradation and elimination. The co-application of low-dose ritonavir as an inhibitor of both P-gp and CYP3A to boost anti-retroviral treatment in HIV infection is an example in clinical practice [113].

CHARACTERIZATION OF P-GP SUBSTRATES

Ideally, CNS drugs should not be P-gp substrates because higher peripheral doses would be required to reach effective concentrations in the CNS. This, in turn, narrows the therapeutic window by increasing the potential for peripheral toxicity [65]. Defining the characteristics of P-gp substrates is therefore of great importance for drug development, especially in view of its role in multidrug resistance.

Recent studies indicate at least four distinct drug binding sites (DBSs) which can switch to high or low affinity configurations for substrates and inhibitors. As the P-gp subunits are mobile within the membrane, so are the DBSs, and they can therefore serve as physically distinct sites or contribute to a larger binding pocket. This can account for P-gp's broad substrate specificity [114] but also for the difficulties in singling out common structural features.

In 1988, Zamora et al. described the structural properties of P-gp substrates roughly as lipid soluble at physiological pH, planar, and carrying a cationic charge [115]. These results have been confirmed and augmented with hydrogen bonding potential, presence of an amine or aromatic rings, and several less

specific features such as molecular mass, surface area, and dimensions [98, 111, 116, 117]. These rather unspecific determinants come as no surprise given P-gp's promiscuity.

3.3.1.4. Breast Cancer Resistance Protein (BCRP)

Breast cancer resistance protein (BCRP, ABCP, MXR) is the product of the ABCG2 gene, located on chromosome 4q22 and spanning > 66 kb. With reference to other (two-dimer) members of the ABC family of transporters, which consist of twelve transmembrane domains (TMD) and two ATP binding domains (ABD), BCRP has been characterized as a half-transporter, as it only has six TMDs and one ABD [118]. BCRP is expressed in the epithelium of the small intestine and small biliary ducts as well as in the breast and placenta [119]. Just as P-gp, it is considered to maintain functional barriers and a limiting factor in drug absorption and distribution, e.g. in topotecan, doxorubicin, and mitoxantrone [120, 121].

4. Materials and methods

4.1. Cheminformatics methods

4.1.1. The chemical graph

In chemical graph theory, the atomic structure of a compound is described as a mathematical graph G whose set of vertices V is a list of the compound's atoms and whose set of edges E corresponds to the bonds between the atoms (vertices) so that

$$G = [V, E]$$

Applied to molecules, this yields a labeled, undirected multigraph which may or may not be disconnected. [18] Usually, hydrogen-depleted graphs are used, i.e. graphs where vertices that represent implicit [19] hydrogen atoms are omitted (Figure 8) unless they are part of a functional group (e.g. –OH or $–NH_2$).

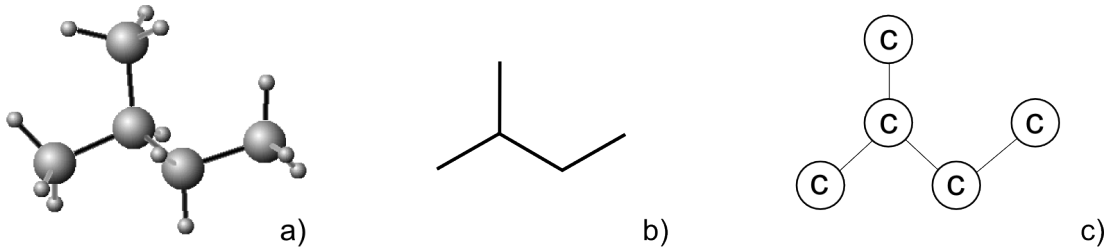


Figure 8 – Isopentane structure with implicit hydrogen atoms (a), its structural formula (b) without implicit hydrogens, and the corresponding molecular graph (c)

The chemical graph is a prime example of Eugene Wigner's 'unreasonable effectiveness of mathematics in the natural sciences' [20], an observation further underlined by the co-evolution that the mathematical field

[18] A graph is connected if a path exists between every pair of vertices. Disconnected graphs do not have this property and therefore consist of two or more components. As a chemical example, one might consider salts (which have a cationic and an anionic component).

[19] Implicit hydrogen atoms are those that fill up valencies in the carbon skeleton.

[20] In his 1960 paper [122], Wigner describes how mathematical models applied to problems within the natural sciences pave the way to a deeper understanding of the problem itself.

of graph theory and the chemical concept of graphs have undergone. [21] Graph theory provides the tools to assess identity between two molecular graphs (graph isomorphism), [22] measures of connectivity (see below), and properties of mathematical graphs have chemical interpretations.

4.1.2. Simple Molecular Input Line Specification (SMILES)

The Simple Molecular Input Line Specification (SMILES) is a language that can be used to unambiguously describe molecular structures as ASCII strings [125]. These are obtained by generating the chemical graph of a molecule (see above) and then printing out a depth-first traversal of the resulting tree [126]. Usually, hydrogen-depleted graphs are used to save computer memory and simplify the final string.

Due to the unambiguousness of SMILES strings, one of their main applications is as a means of generating index keys in chemical databases (i.e. serving as the primary key). Furthermore, the format is used in cheminformatics systems for the calculation of molecular properties (descriptors). Also, because of efforts in the standardization of the notation, anyone who uses the canonical SMILES notation will come up with the same string for a given molecule, thereby easing the exchange of chemical information. [23] Lastly, SMILES provides additional grammar for chemical reactions.

While a given SMILES string corresponds to one and only one molecule (or compound), the reverse does not hold true. Ethanol (C_2H_5O), for example, can be encoded as `OCC`, `[CH3][CH2][OH]`, `C-C-O`, and so on, whereas the canonical SMILES version is `CCO`. However, most current cheminformatics systems and structure editors will arrive at the same representation.

Five general rules govern the generation of SMILES codes:

ATOMS

Atoms are specified by their atomic symbols and enclosed in square brackets ([ ]) unless they are part of the organic subset (C, N, O, P, B, S, F, Cl, Br, and I) or have valences other than the standard ones. Aromatic carbon atoms are given as ‘c’ (small C, Table 4).

---

[21] In fact, when the term ‘graph’ was introduced into mathematics by the English mathematician James J. Sylvester in 1878, he had derived it from the chemical term ‘graphical notation’ for the depiction of a molecule [123]. Sylvester is also credited with coining the mathematical term ‘matrix’.

[22] Isomorphism is assessed by showing that a bijection $f : V(G) \leftrightarrow V(H)$ exists for two compounds G and H so that any two adjacent vertices u and v in the first graph are also adjacent in the second graph [124].

[23] One of the most widely used repositories for chemical structures in SMILES and other formats is the National Center for Biotechnology Information’s (NCBI) PubChem Project. A script for retrieving structural information from PubChem is given in the appendix.

SMILES	Common name	Structure
C	Methane	CH_4
O	Water	H_2O
S	Hydrogen sulfide	H_2S
[S]	Elemental sulfur	S
[H+]	Proton	H^+

Table 3 – Examples of atomic SMILES notation

BONDS

Adjacent atoms are considered to be connected either by a single or an aromatic bond, depending on context. Bonds may be explicitly stated as -, =, and # for single, double, and triple bonds, whereas aromatic bonds are denoted by a colon (":").

SMILES	Common name	Structure
CC	Ethane	CH_3CH_3
C=C	Ethane	$CH_2=CH_2$
C#N	Hydrogen cyanide	HCN
CCO	Ethanol	CH_3CH_2OH
c1ccccc1	Benzene	Benzene ring

Table 4 – Examples of SMILES bond notation

BRANCHES

The branching structure is enclosed in parentheses and assumed to be connected to the left.

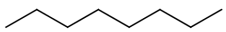
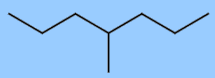
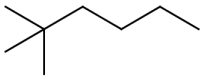
SMILES	Common name	Structure
CCCCCCCC	n-octane	
CCCC(C)CCC	4-methylheptane	
C(CCC(C)(C)C)C	2,2-dimethylhexane	

Table 5 – Examples of SMILES branching notation

CYCLIC STRUCTURES

Because the SMILES notation is non-cyclic, ring structures need to be split. This is done by designating an arbitrary break point in a connected structure and listing the atoms and bond connections as a non-cyclic structure.

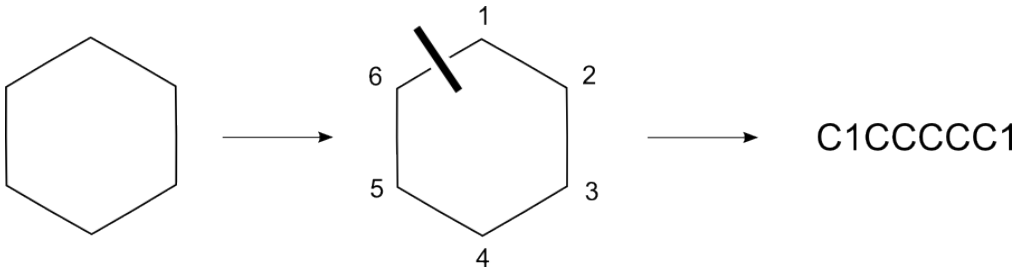


Figure 9 – Example of SMILES notation of a cyclic structure (cyclohexane)

DISCONNECTED FRAGMENTS

Compounds are not necessarily a single connected molecule. Some, for example, are formulated as salts or have counterions. In SMILES notation, this is expressed by joining fragments by a period (“.”). Common salt (*NaCl*), for example, translates to `[Na+].[Cl-]` in SMILES notation.

4.1.3. Canonical and isomeric SMILES

Isomerisms need not be explicitly specified to produce valid SMILES strings. This is arguably a flexibility feature of the standard, allowing for simpler handling of molecules where these aspects are either of no importance or unknown. Stereochemical information can, however, be added to canonical SMILES strings to produce isomeric SMILES whenever the canonical version does not yield the desired results.

For example, the string `[H]C(O)(Br)CC` represents Molecule 1 in Figure 10. The molecule is structurally the same as Molecule 2, specified by the fully-featured isomeric version `[H][C@@](O)(Br)CC`, because it is implicitly assumed that attached substructures are listed in clock-wise order. Molecule 3 is an enantiomer of the first two and is described with `[H][C@](O)(Br)CC`.

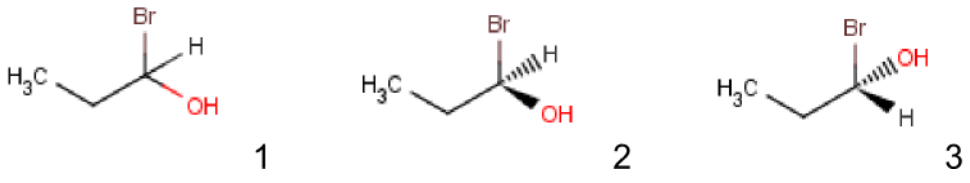


Figure 10 – Tetrahedral chirality in SMILES codes. The structures translate to `[H]C(O)(Br)CC` (Molecule 1), `[H][C@@](O)(Br)CC` (Molecule 2), and `[H][C@](O)(Br)CC` (Molecule 3).

Various other forms of chirality are handled, including allene-like (E and Z symmetry), octahedral, square-planar, and trigonal-pyramidal. Whenever available (and sensible), the isomeric form was used in the databases presented here.

4.1.4. SMILES Arbitrary Target Specification (SMARTS)

A common task in cheminformatics is substructure search. While one might assume this can be achieved by matching character patterns within a SMILES encoded molecule set, this is not the case because a single molecule can be represented by different SMILES strings (see above) and patterns that span branches are hard to retrieve with a character based approach. For instance, a character-based search for the benzene ring (`c1ccccc1`) of the amino acid tryptophan (`N[C@@H](CC1=C[N](C2=C1C=CC=C2)[H])C(O)=O`) is prone to fail.

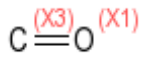
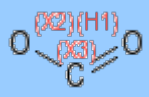
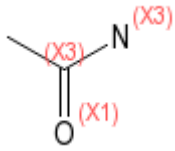
SMARTS	Common name	Target
`[CX3]=[OX1]`	Carbonyl group (low specificity, hits carboxylic acid, esters, ketones, and others)	
`[CX3](=O)[OX2H1]`	Carboxylic acid	
`[NX3][CX3](=[OX1])[#6]`	Amide	

Table 6 – Examples of SMARTS targets, their common names, and their structural representation

SMARTS addresses this problem by providing a language built on the SMILES language. The similarity between the two is so high that almost all SMILES specifications are valid patterns (or rather 'targets') in SMARTS. The necessary flexibility of a pattern matching language is achieved by providing logical operators such as AND and OR, connectivity descriptors (`CX4` hits carbon connected to four other atoms, `CD4` hits quaternary carbon), and cyclicity descriptors. [24] It is also possibly to search with varying degrees of accuracy, e.g. when looking for carbonyl groups (C=O) one might specify a target that hits only carboxylic acid or one that will also find aldehydes, esters, ketones, etc. (see Table 6)

The original development of SMARTS was done by Daylight Systems (who also established the SMILES standard) and their Daylight Theory Manual [127] provides thorough documentation of the standard. Modifications also exist, most notably by OpenEye Scientific Software. This work uses Daylight's

[24] Cyclicity is considered on the basis of the smallest subset of rings (SSSR).

standards. The versatility of the specification has resulted in a wide acceptance. SMARTS functionality can be added to common relational database software, thereby allowing high-throughput screenings of virtual libraries.

4.1.5. Computational complexity of substructure searches

While the notations discussed in the previous section provide a straightforward way to implement a compound library and search for targets, it is important to understand that substructure searches in general are computationally expensive [128]. Finding a substructure in a mathematical graph, which is essentially the same as searching for a target in the molecular graph, has been shown to be an NP-complete (non-polynomial) [25] problem [129].

4.1.6. Chemical similarity

The question of whether or not two structures are similar is not only encountered in QSAR models (Similar Property Principle, see above). It is also of relevance in queries of chemical databases (e.g., 'find ten substances similar to this one') and their characterization (e.g., 'how heterogeneous is the compound library?'). [26]

4.1.6.1. Fingerprinting

Fingerprints are an abstract representation of compounds used in chemical similarity studies and to enhance substructure searches. For a compound *A*, *n* attributes are recorded and stored in a vector V_A of the form

$$V_A = \{X_{1A}, X_{2A}, ..., X_{nA}\}$$

with length n [130]. Attributes may be numerical or binary (indicating absence (0 or OFF) or presence (1 or ON), e.g. of substructures, within the compound) and may be chosen as required by the underlying question (Figure 11). Binary fingerprints are of course preferable when only absence or presence of a feature is evaluated. Numerical descriptors such as molecular weight can be handled either by storing them directly in a field of the fingerprint or by binning [27] them. When examining, for example, the number of hydrogen bond donors, one might define multiple intervals such as 0, (0,2], (2, 4], and ≥ 4.

---

[25] Solutions to these problems can be verified quickly. However, there is no known fast way to come up with a solution in the first place.

[26] A point worth noting, esp. when dealing with similarity measures (see below), is that perfect *similarity* does not imply *identity*. Similarity is an approximation and approaches identity as more and more features are taken into consideration.

[27] In binning, a continuous numerical range is separated into a small number of discrete segments (bins). These bins are assigned integer numbers in ascending range. In this way, continuous variables are discretized.

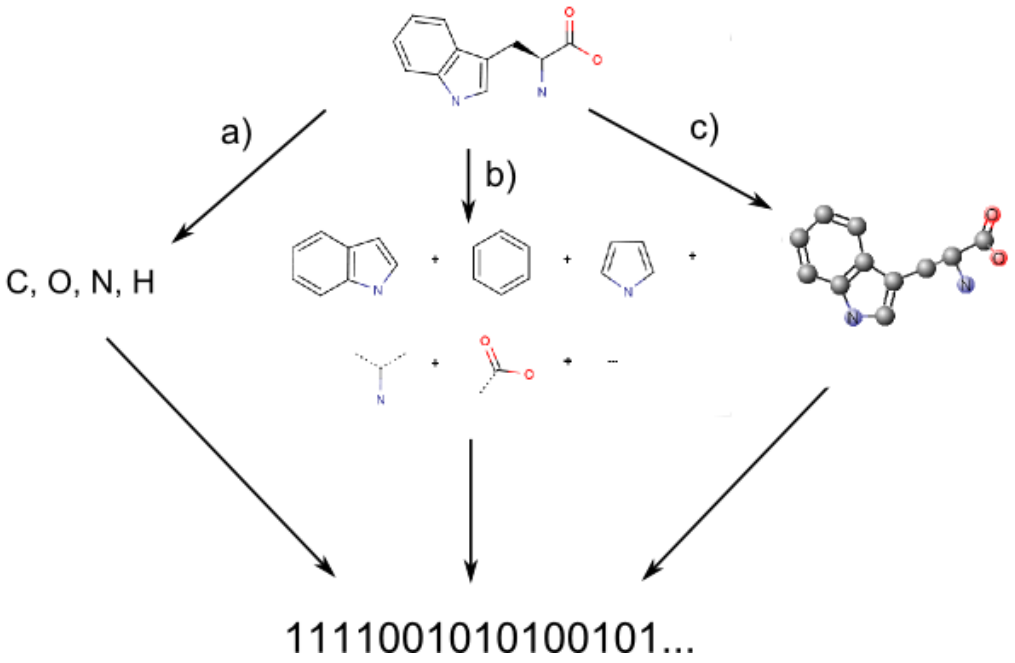


Figure 11 – Example of a binary fingerprint of lysine based on the elemental composition (a), interesting substructures (b), and spatial dimensions (c). The exact composition of fingerprints varies with compound libraries and purpose.

Fingerprints are stored alongside the compound. Queries can then be designed to make use of these strings (e.g. if the desired target is an indole ring (2,3-benzopyrrole, C_8H_7N), the corresponding query could first check for the presence of nitrogen or for at least two ring systems).

4.1.6.2. Assessment of chemical similarity

While searches of substructures are probably the most common queries in chemical databases, similarity based queries are gaining in prevalence. They can refine a substructure search by specifying a context to filter potential hits (i.e. 'look for phenol rings in structures like this.'). Also, given a molecule of interest (bait), queries can be formed that return similar compounds (i.e. 'give me ten compounds like this.').

There is no accepted definition of what chemical similarity or diversity is, even though one generally recognizes it (or, more often, its absence). [28] Consequently, there is no agreed-upon way of ranking molecules by their similarity [131, 132]. This becomes clear when one looks at the different coefficients (Table 7) that are currently in use for measuring similarity or distance in the chemical space.

[28] In fact, many chemical definitions are exclusive: carbocycles are defined as not containing heteroatoms, aromatics as lacking aliphatic components, and so on.

Coefficient	Expression
Tanimoto coefficient (Jaccard) [29]	$\frac{c}{a+b-c}$
Cosine coefficient (Ochiai coefficient)	$\frac{c}{\sqrt{a+b}}$
Hamming distance (Manhattan distance, city-block distance)	$a+b-2c$
Euclidean distance [30]	$(a+b-2c)^{\frac{1}{2}}$
Soergel distance	$\frac{1-c}{a+b-c} = \frac{a+b-2c}{a+b-c}$
Russel-Rao coefficient	$\frac{c}{m}$
Forbes coefficient	$\frac{cm}{ab}$

Table 7 – Similarity coefficients (synonymous names) for use in binary fingerprints, where *a* and *b* are the bits set to 1 in binary fingerprints of length *m* bits of two molecules A and B. *c* is the number of bits set in both *a* and *b* (logical AND; adapted from [130, 131])

All of these expressions are closely related and most may be interpreted geometrically as distances in a hyper-dimensional space spanned by the available attributes. Depending on the data set considered, e.g. the size distribution, some measures may fare better than others, esp. when it comes to judging a substitution series of smaller molecules. The Tanimoto coefficient is the most widespread [131] and also the one used in this work (an implementation is given in the appendix). The example in Figure 12 illustrates one of the advantages of this measure [31] over the also quite popular Hamming distance. Identical substitutions (in this case –O-methyl for –Cl) tend to be overrated by measures that do not normalize like Tanimoto's equation.

[29] Tanimoto's coefficient builds on Jaccard's coefficient, a widely used similarity measure for binary data. The coefficient is therefore also known as 'Extended Jaccard', is applicable to real numbers, and, when applied to binary values, the same as Jaccard's coefficient.

[30] For binary keys, this equals the square root of the Tanimoto index [133].

[31] In the example, the complementary of Tanimoto's coefficient, Soergel distance, is used. For similarity indices, the complementary (1 – measure) corresponds to distance.

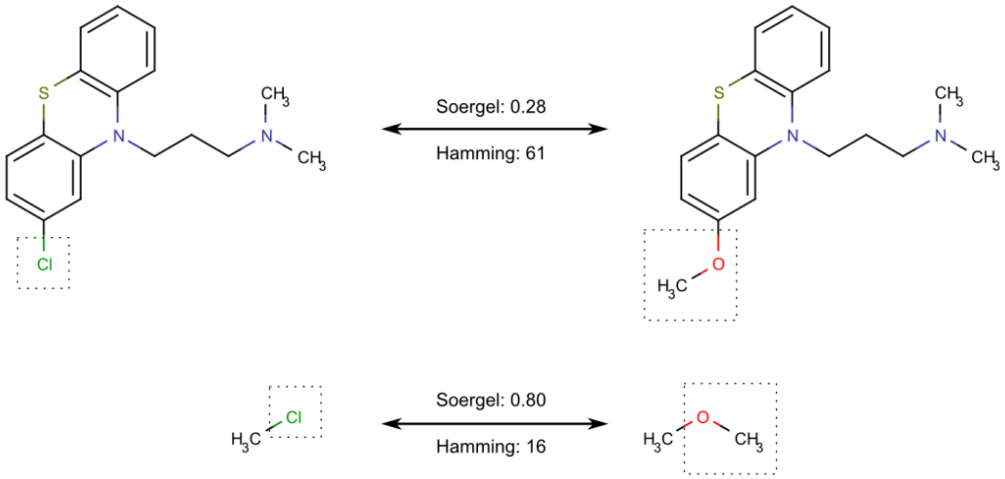


Figure 12 – Soergel and Hamming distance for two pairs of molecules with identical substitutions (adapted from [134])

4.1.6.3. Applicability and interpretation of fingerprints

An expert in the domain can quickly scan a chemical structure and extract important properties by noting absence or presence of certain functions or properties. Fingerprints seek to automate this by evaluating a given set of features for each compound. Feature selection may be general (in order to compare different heterogeneous databases) or more specific to a domain (in order to single out interesting leads). As a quality measure, chemical similarity is a valuable parameter that describes how diverse the investigated set is. Furthermore, it reflects the quality of the choice of descriptors.

4.1.7. Descriptors

The molecular structure is not easily accessible to numerical analysis and learning methods. It is, however, possible to derive parameters (descriptors) from molecules that describe their physicochemical properties, electronic features, and so on. For use in QSAR studies, they should be easy to interpret and calculate as well as sensitive to small variations in the properties they measure [135]. The following sections give an overview of the different classes of calculations commonly employed.

4.1.7.1. Elemental analysis

Calculations of this class are 1D QSAR descriptors as they are computed directly from the sum formula (i.e. C_2H_6O for ethanol).

MASS CALCULATIONS

One of the most straightforward ways of characterizing a compound is its mass. This can either be found by summing up the atomic weight of the constituents (mass, for ethanol: 46.0864) or the weights of the most common isotope found in nature (exact mass, for ethanol: 46.0419).

COMPOSITION COUNTS

The total atom count and the count of heteroatoms (all elements except for carbon and hydrogen) are of biochemical relevance. Most cheminformatics systems also analyze isotope composition although its importance in biochemical models is very low because biological systems are generally insensitive to isotopes [136]. [32]

APPLICABILITY AND INTERPRETABILITY

Molecular mass, of course, is an important feature of compounds, and appears in very simple but effective models [26]. Elemental analysis is also useful in identifying outliers and pre-screening of databases (e.g. removing heavy structures in a search for small drug-like compounds).

4.1.7.2. Constitutional analysis

Constitutional descriptors are slightly more abstract than their elemental counterparts and require the structural formula. Among the simplest ones are total counts of atoms and (single, double, aromatic, etc.) bonds. The number of general classes of substructures may also be counted (aliphatic, aromatic, and hetero rings) as well as specific motifs (e.g. phenol groups, amino acid residues). Special care must be taken in practice to precisely formulate such queries. For example, the explosive 2,4,6-trinitrotoluene (TNT) has strong radial symmetry. If the count of nitrobenzene groups has to be determined, a SMARTS query such as `c1ccccc1-[#7+](=[#8])-[#8-]` will return three hits (Figure 13).

A less trivial constitutional descriptor is the cyclomatic number. It gives the number of independent cycles (rings) in a molecule and is equal to the minimum number of edges that need to be removed in order to transform it to an acyclic graph [138]. This must not be confused with the term cyclicity from graph theory, which states the total number of cycles (e.g. including cycles containing other cycles).

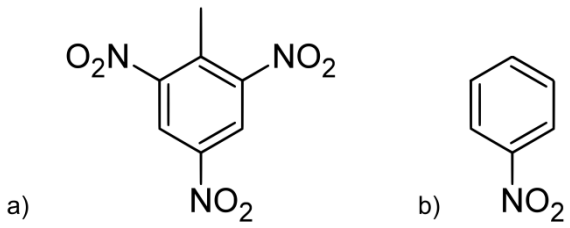


Figure 13 – Substructure matching. In 2,4,6-trinitrotoluene (TNT, a), the substructure motif of nitrobenzene (b) can be found three times.

[32] Radio-assays exploit this characteristic of enzyme kinetics by introducing radioactively marked substrates which are metabolized to the same degree as unmarked substances. Sometimes, however, a rate-determining step is the breaking or formation of hydrogen bonds, which is slowed down by replacing normal hydrogen (^{1}H) by deuterium (^{2}H), the bonds of which are harder to break [137].

APPLICABILITY AND INTERPRETABILITY

The descriptors outlined above disregard molecular geometry. They do not change with conformation, do not distinguish isomers, are among the most commonly used descriptors, and lend themselves to fingerprinting (see above). Interpretation is intuitive, e.g. presence of many aromatic bonds and features suggests lipophilicity, and so on.

4.1.7.3. Electronic descriptors and charge analysis

Electro-chemical properties are central to chemical reactions and biological activity, and charged protein domains and membranes greatly influence ADME behavior. The following survey of descriptors ranges from simple charge counts to rich properties encompassing molecular shape and size.

ATOMIC CHARGE

Maximum and total positive and negative charges, total absolute charges, and analysis of electrical dipole moments for atoms and molecules are among the most popular examples.

HYDROGEN-BONDING DESCRIPTORS (HBA, HBD)

Hydrogen bonds are weak bonds (5 – 30 kJ/mol) between an electronegative atom and a hydrogen atom connected to another electronegative atom. It is a dipole-dipole interaction and should not be confused with covalent bonds [139]. This type of bond is encountered both within and between molecules, and in organic (e.g. DNA) as well as in inorganic (e.g. water [33]) compounds.

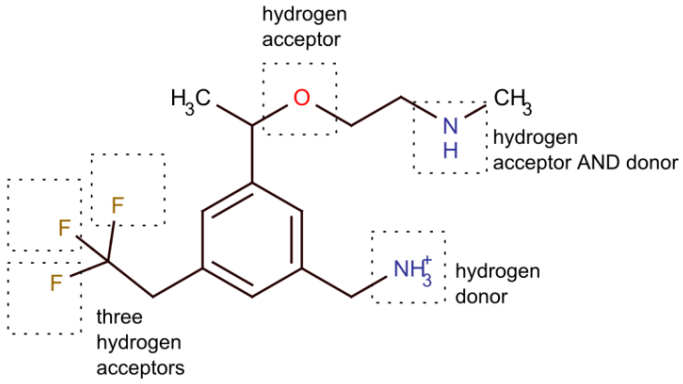


Figure 14 – Hydrogen donor and acceptor sites in a hypothetical molecule. Donor count (HBD) is 2 and acceptor count (HBA) is 4 with 4 donor sites and 5 acceptor sites.

[33] This is where hydrogen bonds were first analyzed. For a molecule of low mass (~ 18 g/mol), water has unexpectedly high boiling and melting points, which can be explained by their tendency to form long, hydrogen-bond stabilized chains.

Functional groups can be hydrogen bond donors (HBD), acceptors (HBA) [34] or both (i.e. amphiprotic [140]), as illustrated in Figure 14. HBDs have electron-withdrawing substituents (e.g. –OH, –SH, –FH) while HBAs have electron-donating substituents (e.g. –O, –S, –F, –PO) [138]. Consequently, most computational methods of calculating these descriptors are based on group contribution (e.g. [141]). Hydrogen-bonding characteristics also depend on the environment, as the protonization of functional groups changes (Figure 15).

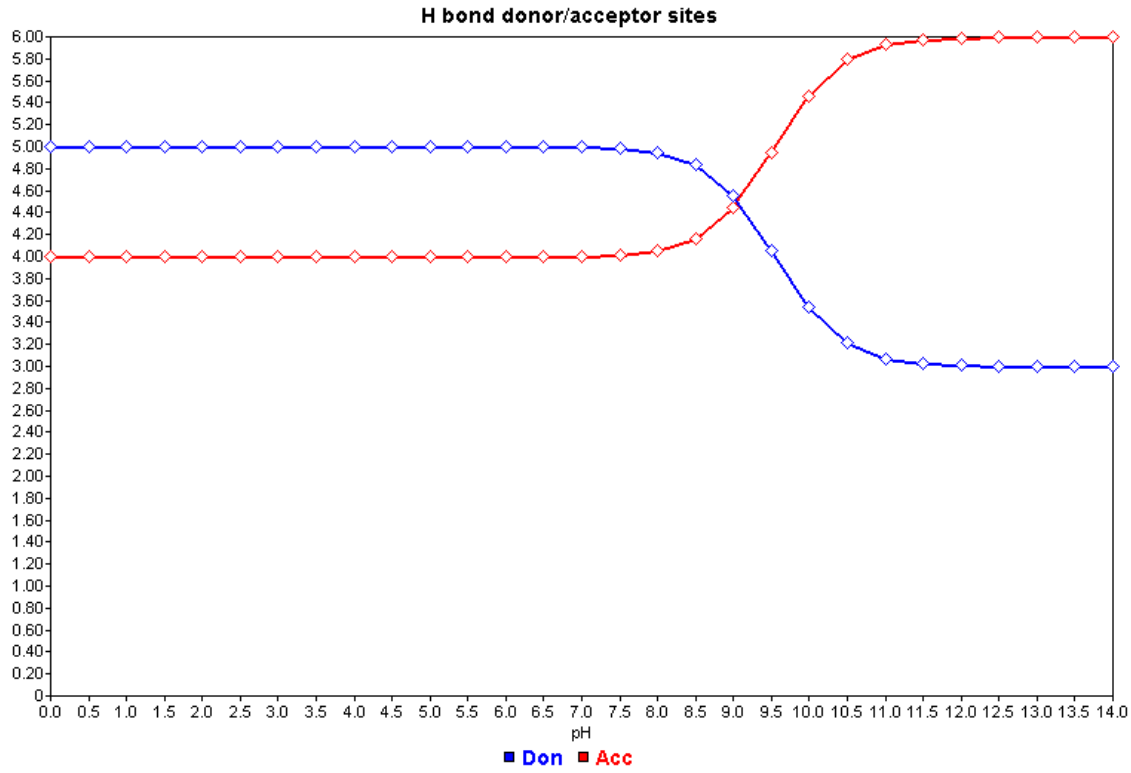


Figure 15 – Hydrogen bond donor and acceptor sites (ordinate) with increasing pH (abscissa) for the hypothetical compound given in *Figure 14*.

VAN DER WAALS SURFACE AND CHARGED PARTIAL SURFACE AREA DESCRIPTORS (CPSA)

When compounds are immersed in a medium, e.g. drugs during oral absorption or in different compartments such as plasma or CSF, non-covalent interactions take place between the compound and its surroundings. The most simple of those is solution (or suspension) in water. For these situations, the

[34] Count of the functional groups with said properties should not be confused with the number of HBD and HBA *sites*, where the number of donatable or acceptable hydrogens is counted.

van der Waals surface area (VDWSA) is of great utility. This views molecules as sets of overlapping hard spheres [35] with a size determined by the van der Waals radii of their atoms (Figure 16a).

In 1990, Stanton and Jurs introduced a set of descriptors that link charge to shape [144] by building on work by Lee and Richards who refined the VDWSA [145]. Here, the same three-dimensional representation is traced with another sphere, this time representing a solvent molecule (e.g. water). The basis for calculations is a contact surface defined by the center point of the tracing sphere, which is complemented by attractive and repulsive forces over each point as defined by the electron distribution of the original molecule (Figure 16). This contact surface is what defines polar interactions.

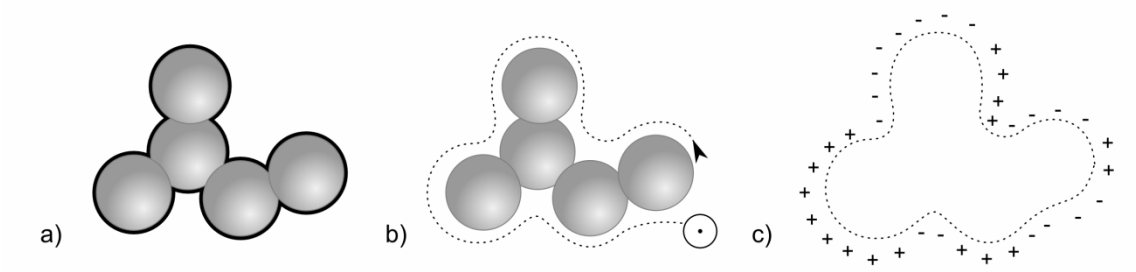


Figure 16 – Molecular representation used in charged partial surface area (CPSA) calculations. Atoms are represented as solid overlapping spheres with a radius corresponding to the van der Waals radius (solid black line) of each atom (a). For CPSA, this is traced with another sphere (⊙) representing a solvent (b). The traced surface (c) is shown as a dotted line (adapted from [134, 144]).

The most basic values give the partial positive and negative surface areas. Weighing the total positive and negative charges of the molecule against partial charges and individual atoms (and several other correlations) results in a total of 25 different descriptors. Stanton and Jurs successfully linked their descriptors to boiling points and surface tension.

TOPOLOGICAL POLAR SURFACE AREA (TPSA)

Polar surface area (PSA) descriptors such as CPSA quickly proved successful in predicting membrane transport and drug properties. Their calculation, however, is time consuming because a 3D-model has to be generated for every compound. In 2000, Ertl described a fragmentation-based scheme based on a study of over 30'000 substances [146]. He gives 41 fragments and their contribution to polar surface area. Using this approach, PSA can be estimated by summing up individual contributions. Of course, as

[35] This is highly reminiscent of the ball-and-stick molecular model in John Dalton's atomic theory. While this has been largely abandoned, the paradigm works well in certain areas of molecular modeling such as this. Of course, restrictions apply [142, 143], e.g. when conformational changes expose surfaces during ligand-receptor interactions.

molecules increase in size, polar groups that contribute to TPSA may become inaccessible. Therefore, this descriptor does not do well for macromolecules. [36]

Descriptor	Meaning	Equation
PPSA-1	Partial Positive Surface Area	$\Sigma(+SA_i)$
PPSA-2	Total Charge Weighted PPSA	$(\Sigma(+SA_i))\ Q^+_T$
PPSA-3	Atomic Charge Weighted PPSA	$(\Sigma(+SA_i))\ Q^+_i$
PNSA-1	Partial Negative Surface Area	$\Sigma(-SA_i)$
PNSA-2	Total Charge Weighted PNSA	$(\Sigma(-SA_i))\ Q^-_T$
PNSA-3	Atomic Charge Weighted PNSA	$(\Sigma(-SA_i))\ Q^-_i$
DPSA-1, DPSA-2, DPSA-3	Difference in Charged Partial Surface Areas	(PPSA-1 – PNSA-1), (PPSA-2 – PNSA2), (PPSA-3 – PNSA-3)
FPSA-1, FNSA-1, FPSA-2, FNSA-2, FPSA-3, FNSA-3	Fractional Charged Partial Surface Areas	CPSA / total molecular surface area
WPSA-1, WNSA-1, WPSA-2, WNSA-2, WPSA-3, WNSA-3	Surface Weighted Charged Partial Surface Areas	(CPSA)(total molecular surface area) / 1000
RPCG	Relative Positive Charge	(charge of most positive atom) / (sum total positive charge)
RNCG	Relative Negative Charge	(charge of most negative atom) / (sum total negative charge)
RPCS	Relative Positive Charged Surface Area	$(SA_{MPOS})(RPCG)$
RNCS	Relative Negative Charged Surface Area	$(SA_{MNEG})(RNCG)$

Table 8 – Charged Partial Surface Area (CPSA) descriptors [144].

APPLICABILITY AND INTERPRETABILITY

As was seen before, electronic charge influences intra- and inter-molecular attraction. Hydrogen bonding is of direct relevance to many physicochemical properties such as compressibility, boiling and melting points, and molar refractivity. Along with partition coefficients (see below), these descriptors are good

[36] The original study was limited to molecular mass < 800 and > 100.

predictors of passive membrane permeability and absorption [147]. PSA descriptors deliver good results in these areas as well [148-151] and the special usefulness in combination with connectivity indices for biological studies has been shown [152].

Although polar interaction plays an important part in transport phenomena and the good explanatory power of such descriptors therefore does not come as a surprise, one should remember that some of the PSA descriptors were developed using drug databases (e.g. the World Drug Index for TPSA [146]). This preference for drug molecules should be kept in mind when interpreting polar surface contributions.

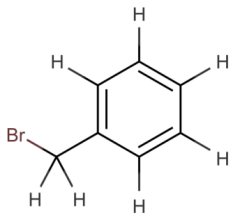
4.1.7.4. Partitioning

PARTITION COEFFICIENT (LOGP)

The hydrophobic properties of compounds have been found to correlate well with their pharmacokinetic and pharmacodynamic behavior [18, 153]. As a measure of hydrophobicity, the partitioning of the unionized compound in a two-phase system with n-octanol (hydrophobic compartment) and water (hydrophilic compartment) may be determined experimentally [37] (see 3.1.2.2). The logarithm of the solute's distribution ratio within these compartments gives *logP*:

$$\log P_{oct/wat} = \log\left(\frac{[solute]_{oct}}{[solute]_{wat}}\right)$$

Several computational approximations for this descriptor exist. First attempts were made in 1964 with substitution studies in aromatic rings by Fujita, Iwasa, and Hansch [154], who gave π values for over 60 substituents. Extrapolation in other series eventually led to the development of CLOGP [155] which is still in use today. CLOGP splits molecules at isolating carbons, meaning C atoms without double or triple bonds to heteroatoms. These carbons and their connected hydrogen atoms are considered hydrophobic, and the remaining contribute to the molecules polarity. LogP values are computed by summing up these contributions with appropriate correction factors (Figure 17).



Bromide	0.480
1 aliphatic isolating carbons	0.195
6 aromatic isolating carbons	0.780
7 H on isolating carbons	1.589
1 chain bond	- 0.120
TOTAL	2.924

Figure 17 – Example of a CLOGP calculation of benzyl bromide (adapted from [134])

[37] Experimental values may be difficult to obtain when dealing with very lipophilic or zwitterionic compounds [134].

Atom-centric approaches, such as ALOGP [156] and XLOGP [157], take into account not only the atom type but also its environment [158][38]. Of course, it is not always clear how to properly fragment a (large) molecule, and for some fragments, these contributions are not known (although systems like CLOGP can provide estimates).

DISTRIBUTION COEFFICIENT (LOGD)

An adaptation of logP for ionizable compounds is the distribution coefficient (logD). The experimental setup is analogous to that of the partition coefficient [139] with the exception that the aqueous phase is buffered to a certain pH (so that it is not altered by introduction of the compound). Both the ionized as well as the neutral form of the solute are measured to give *logD* depending on pH [39] (Figure 18):

$$\log D_{oct/wat} = \log\left(\frac{[solute]_{oct}}{[solute]_{wat}^{neutral} + [solute]_{wat}^{ionized}}\right)$$

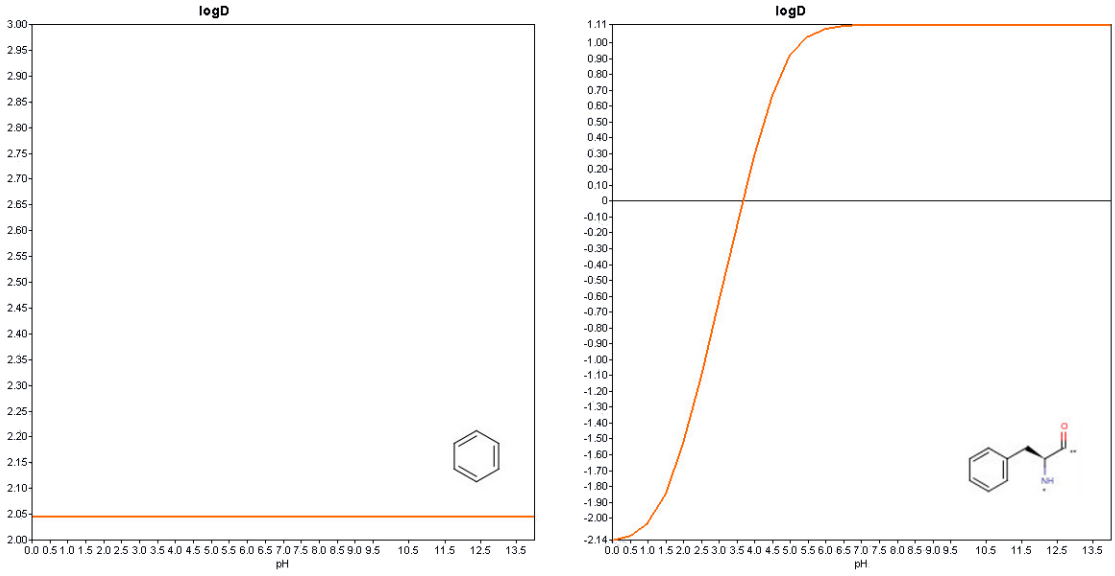


Figure 18 – *logD* of benzene (*logP* = 2.05, left) and phenylalanine (*logP* = 1.11, right) over the entire pH range (0 to 14)

With this equation, distribution and hydrophobic character can be estimated for given environments such as the blood plasma (pH = 7.4), the stomach (acidic, with a fasted pH of about 1.4 – 2.1 and a fed pH of

[38] CLOGP and ALOGP, both widespread, are at two ends of the spectrum of computational logP prediction. CLOGP uses longer fragments and heavy correction for intramolecular forces, whereas ALOGP uses atom-size fragments and does not perform these corrections [159].

[39] It follows for un-ionizable compounds that *logD* = *logP*. For ionizable compounds, *logP* = *logD(7.0)*.

3.0 - 7.0 [160]), and so on. State of ionization therefore has major implications for solubility, permeability, and, finally, absorption.

APPLICABILITY AND INTERPRETABILITY

As stated before, a compound's hydrophobic character is of importance in pharmacological studies. Hydrophilic compounds tend to be found in hydrophilic compartments such as the blood plasma whereas lipophilic structures distribute to lipid bilayers and the like [147]. Even though *logP* plays a major role in many pharmacological models, its very definition rests on the distribution of an *unionized* solute. Most drugs, however, are ionizable, [40] and hence *logD* would seem more appropriate, especially when absorption needs to be considered.

The number of different software systems for the prediction of logP values suggests that none is truly superior. In fact, some studies show poor correlation (R^2 in unseen compounds for ALOGP of 0.75 and 0.72 for CLOGP) of computed logP values with observations [28]. Choosing one for modeling depends on cost, hardware considerations, and simply how well it performs for the endpoint.

4.1.7.5. Molecular connectivity indices

In a series of isomers, physicochemical parameters, e.g. boiling points, vary with the shape or the degree of branching (Figure 19). The class of molecular connectivity indices formalizes the rather intuitive notion of how complex a structure actually is. In general, these descriptors are graph-theoretical invariants. [41]

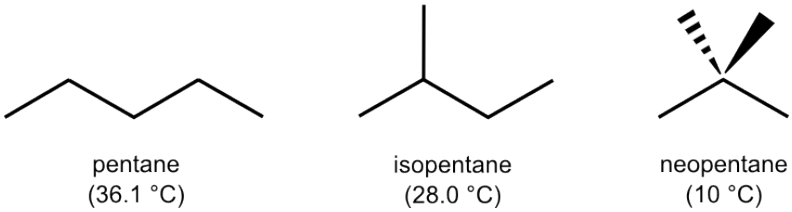


Figure 19 – Structure of pentane isomers with boiling points

[40] This important point is noted in many logP methodology studies where, for example, corrections for the zwitterion character of amino acids are accounted for explicitly by a correction factor.

[41] Invariants in mathematics are properties that remain constant after transformation of the underlying object. Applied to working with labeled multigraphs (such as the chemical graph), invariants yield the same result regardless of how the graph is displayed or in which order it is traversed. As was seen before, SMILES notation allows for countless representations of the same molecule (see Section 4.1.2). Invariants are therefore of special importance for consistent results.

WIENER, PLATT, AND RANDIC INDICES

The oldest of these, the Wiener number (W), is the sum of distances between all pairs of vertices [161]. To compute it, one takes the hydrogen-depleted graph, adds up the length of the shortest path between every pair of atoms, and then multiplies it by factor ½ because the chemical graph is undirected [42] and any path between two atoms will be traversed twice. In the Wiener index (W) equation, *N* is the number of non-hydrogen atoms and d_{ij} is the shortest distance between two atoms *i* and *j*.

$$W = \frac{1}{2}\sum_{i,j}^{N} d_{ij}$$

The Mean Wiener Index ($\overline{W}$) averages these distances. For a graph with n vertices, one arrives at the following equation:

$$\overline{W} = \frac{W}{\binom{n}{2}}$$

Building on this and work by Platt [162] (who details a descriptor that takes neighboring features into account), Randic developed a connectivity index for alkane branching that separates bonds based on the valencies *m* and *n* of the two partners into different classes of *(m, n)*-edges (Figure 20).

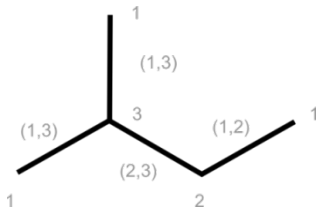


Figure 20 – Partial contributions to the Randic index in isopentane. The numbers near the atom centers denote the valencies and the contributions are given in parentheses along the edges.

The contributive numerical values of the edge classes are added to give the Randic index [163]:

$$\mathrm{X} = \sum (d_i d_j)^{-\frac{1}{2}}$$

[42] i.e. if nodes n_i and n_j are connected, then so are node n_j and n_i. In directed graphs, a node can have predecessors and ancestors, in which case a connection only goes one way.

The individual terms $d_i d_j$ are called 'edge connectivity' and provide a measure of how accessible each bond within the structure is to intermolecular interaction, that is, the higher the edge connectivity, the lower the accessibility [164].

VALENCE CONNECTIVITY INDICES

Kier and Hall refined Randic's index by considering valences and also longer fragments than just one bond [165, 166] in a large set of topological descriptors. The basis for many of these calculations is the simple delta (δ) and the valence delta (δ^v) value for every non-hydrogen atom. Given the number of attached hydrogens (*h*), its sigma electrons (σ) and valence electrons (Z^v), the delta values are defined as [43]

$$\delta = \sigma - h \quad ; \quad \delta^V = Z^v - h$$

For elements beyond fluorine, δ^v is modified as

$$\delta^v = \frac{Z^v - h}{Z - Z^v - 1}$$

where Z is the atomic number of the element. The first set of indices proposed by Kier and Hall are the *chi molecular connectivity indices* (^mX and $^mX^v$), which are sums of δ and δ^v over a bond length *m*. For zero order paths (individual atoms only) in molecules with *n* atoms, this gives

$$^0X = \sum_{i=1}^{n} (\delta_i)^{-\frac{1}{2}} \quad ; \quad ^0X^v = \sum_{i=1}^{n} (\delta_i^v)^{-\frac{1}{2}}$$

Of course, this does not encode any structural information. This can be achieved by summing over bonds, e.g. for bond length *m* = 1 with two adjacent atoms at position *i* and *j*:

$$^1X = \sum_{i=1, j \neq i}^{n} (\delta_i \delta_j)^{-\frac{1}{2}} \quad ; \quad ^1X^v = \sum_{i=1, j \neq i}^{n} (\delta_i^v \delta_j^v)^{-\frac{1}{2}}$$

The *kappa shape indices* [167] provide a further link to shape by comparing bond counts *P* of molecules to extreme shapes (i.e. those that yield maximal or minimal index values) as provided by graph theory.

[43] Thus, while CH_3 differs from $–CH_2–$ in its δ value, it does not differ from $–NH_3$. Valence delta, however, is different.

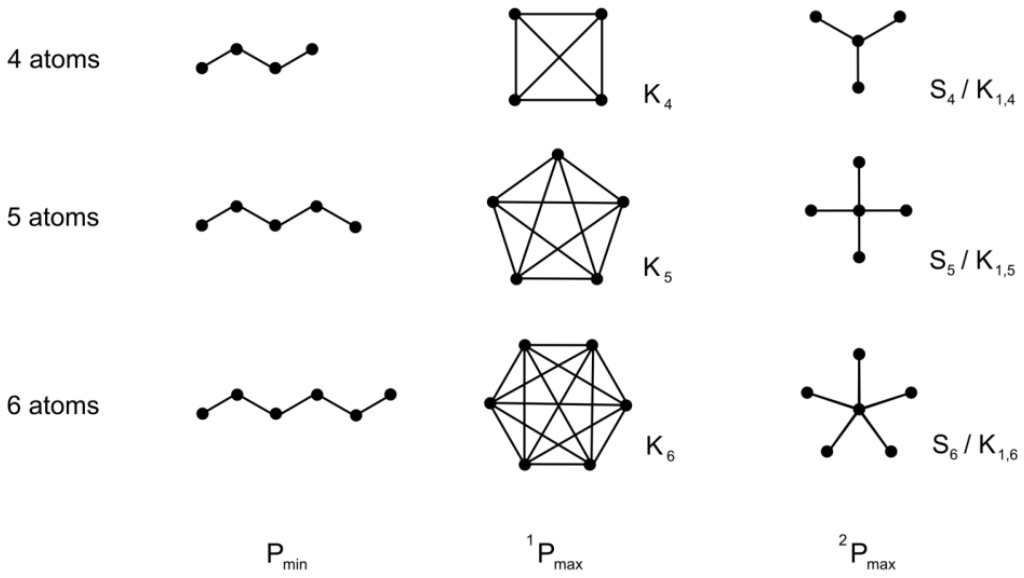


Figure 21 – Extreme shapes in first and second order kappa connectivity indices for 4 – 6 atoms. Minimum values are achieved by linear chains (P_{min}) and maximum values ($^1P_{max}$ and $^2P_{max}$) are produced by complete graphs (adapted from [134]).

ELECTRO-TOPOLOGICAL STATE (E-STATE)

Further work by Kier and Hall uses the intrinsic state of an atom and the electronic and topological contexts (or fields) within which it is embedded in a molecule (electro-topological state (E-state)). The first consideration [20, 168] is the count of sigma- and non-sigma-electrons (as was introduced above as simple and valence delta values δ and δ^v) that encodes an atom's intrinsic value I:

The electron configuration not only correlates with electronic properties such as electronegativity but also to its topology (the degree to which the atom is buried within the structure). The actual E-state (S) of an atom is defined as the sum of *I* and the effects *ΔI* of each of the other atoms *j* (r_{ij} being the distance between atoms *i* and *j*) on the current one:

$$\Delta I = \sum_j \frac{I_i - I_j}{r_{ij}^2}$$

The E-state is affected not only by a change in chain length. Effects are also seen when branches and higher bonds are introduced, mirroring the altered accessibility as atoms are buried. Substitution of a

heteroatom not only changes valencies but electronegativity, too, again changing the E-state [168]. Values for an entire molecule are usually achieved by calculating the mean square value for every atom [134].

SMALLEST ASSOCIATED BINARY LABELS (SABL)

The chemical graph is abstracted to such a degree that it allows the application of other tools from graph theory to molecules. For instance, connectivity might also be characterized by the length of the Smallest Associated Binary Labels (SABL), i.e. the shortest sequence of bits required to define the connections within a molecule as represented by the corresponding adjacency matrix (Figure 22). [44] For pentane (Figure 22b), the SABL is "10, 101, 1010, 10100, 1000" (length = 18 bits), for isopentane (Figure 22d) "1, 1, 10, 101, 11010" (length = 12 bits), and "1, 1, 1, 1, 11110" (length = 9 bits) for neopentane (Figure 22f).

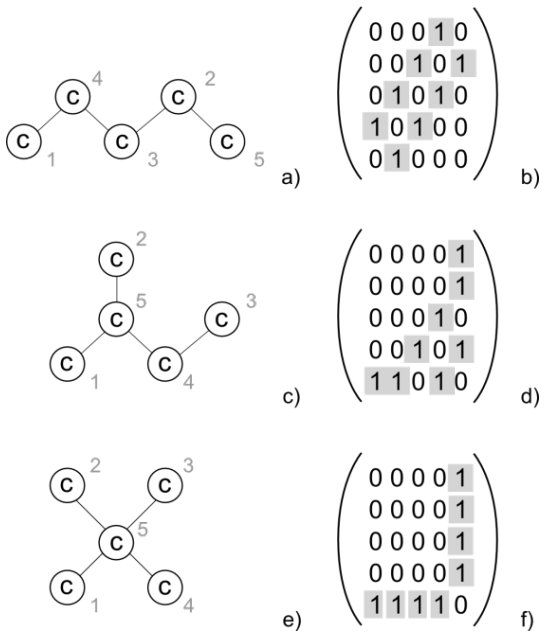


Figure 22 – Hydrogen-suppressed molecular graphs and adjacency matrices for pentane (a, b), isopentane (c, d), and neopentane (e, f).

APPLICABILITY AND INTERPRETABILITY

The distinct advantage of connectivity descriptors is that they can be calculated from the molecular graph itself, require no experimental verification, and correlate well with many physicochemical parameters [163, 165]. However, even though their ease of interpretation suggest generality, very abstract descriptors such as the Wiener index or SABL work best in homogenous compound sets, e.g. in substitution schemes. Newer work addresses this problem and extends well known indices (for instance the Szeged Index [169,

[44] An adjacency matrix has *n x n* dimensions where *n* is the number of nodes and whose elements $a_{i,j}$ are equal to the number of connections between node n_i and n_j (0 for non-adjacent nodes).

170] and Overall Wiener Index [171]). Still, for very heterogeneous sets, information-rich descriptors as detailed by Kier and Hall are preferable.

A general interpretation for connectivity is notoriously hard to give. Edge connectivity as a measure of accessibility of individual bonds (e.g. in the Randic index) can be seen as contribution to interaction between two different molecular species.

4.1.7.6. Structural and geometrical indicators

MOLAR REFRACTIVITY

Several descriptors exist to give numerical approximations of molecular dimensions and shape. The molar refractivity (*MR*) index is a very straight-forward way to do this. It links molecular weight (*MW*) to molecular density (*d*) and refractivity (*n*):

$$MR = \frac{(n^2 - 1)}{(n^2 + 1)} \frac{MW}{d}$$

Because refractivity changes only minimally between different molecular species, MR helps to gain an impression of the steric bulk and thereby complements connectivity indices such as the Wiener or Randic index (see 4.1.7.5).

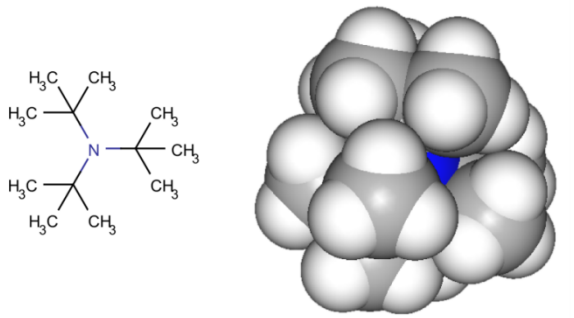


Figure 23 – Steric bulk and accessibility of the lone pair electrons of nitrogen (blue) in tri-tert-butylamine (N,N-ditert-butyl-2-methylpropan-2-amine).

MOMENTS OF INERTIA

Moments of inertia are simple physical indicators of mass distribution and symmetry. They are calculated for each atom *i* of the *n* atoms in a molecule by summing up the product of the mass and the perpendicular distance to every other atom along any of the principal axes of a three-dimensional coordinate system as

$$MOMI = \sum_{i=1}^{n} m_i r^2$$

or, explicitly, for the axes X, Y, and Z, given atomic coordinates as *(x,y,z)*:

$$MOMIX = \sum_{i=1}^{n} m_i (y_i^2 + z_i^2) \quad ; \quad MOMIY = \sum_{i=1}^{n} m_i (x_i^2 + z_i^2) \quad ; \quad MOMIZ = \sum_{i=1}^{n} m_i (x_i^2 + y_i^2)$$

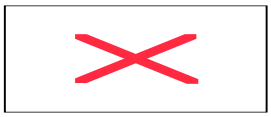
Ratios defined from these values (MOMIXY, MOMIYZ, and MOMIXZ) are also frequently used. A closely related descriptor is the molecular radius of gyration about the center of gravity [172]

$$MOMIR = \sqrt{\frac{\sum_{i=1}^{n} m_i r^2}{MW}}$$

where *MW* is molecular weight. These descriptors not only serve to characterize rotational dynamics but also hint at molecular symmetry. For example, in CCl_4 the principle moments are all equal (*MOMIX* = *MOMIY* = *MOMIZ*), and the molecule can be said to have the symmetry of a spherical top. Similar relationships can be found for linear (e.g. C_2H_2), planar (e.g. C_6H_6), and other symmetries.

GRAVITATIONAL INDICES

Much like the moments of inertia and molecular refractivity, gravitational indices reflect mass distribution and bulk of a molecule by relating measures of mass to measures of intramolecular distance. In its simplest form, calculations are performed for pair of atoms *i* and *j* with mass m_i and m_j and distance r_{ij}:



Similar calculations limit themselves only to bonded atom pairs or only heavy atoms. Square and cubic roots are also frequently employed [138].

APPLICABILITY AND INTERPRETABILITY

The attractiveness of this class of descriptors lies not only in their low computational expense. They are also easily interpreted from a physical point of view. Moments of inertia are, in the end, common concepts in everyday life (consider, for example, how ice skaters reduce them along two axes (linear symmetry) to increase speed in pirouettes, or water divers along three axes (spherical top symmetry) during spins).

4.2. Machine learning methods

4.2.1. Overview of machine learning methodology

Machine learning (ML) is a discipline of artificial intelligence (AI) and deals with the learning and representation of knowledge. Some of its methods are taken from multivariate statistics (e.g. regression), others are distinctively AI (e.g. neural nets), although both fields borrow from each other. Data mining, the task of sifting through large data sets, pattern recognition, and data warehousing, the task of storing and archiving large data sets, are closely linked to ML.

The simplest way of representing knowledge or experience is by listing all previously encountered instances and finding the instances to be predicted in this list. Of course, this would not be considered learning in common understanding, and new events could easily fail to be classified correctly if they do not exactly match entries in the list. The main goal in ML is therefore to derive concepts from existing data, generalize them, and successfully apply them to unseen instances (i.e. predict them). [45]

4.2.1.1. Supervised versus unsupervised learning

Generally, learning can occur supervised or unsupervised. In supervised learning, the list of instances (the data set) as a vector of attributes is augmented by another dimension that denotes the class or outcome of this event (i.e. labeled data). For example, a data set on acute pancreatitis patients would likely contain reported symptoms, laboratory values, and examination results, plus a column ‘outcome’ that states whether or not the patient was actually diagnosed with the condition. A supervised learning method would take a random sample of this data (the training set), build a model based on this, check its validity based on the rest of the data (the test set), and repeat the process until the test set is predicted with good accuracy. Decision trees and Naïve Bayes learning are classical examples.

In unsupervised learning, the method is left to come up with its own classes. These can be validated on other, previously unseen, data. The advantage of this is that a system may be revealed as more complex (or less) than previously thought. Notable methods that use this approach are clustering and Kohonen’s Self-Organizing Maps (SOM).

[45] Memorization is an extreme form of what is known as ‘lazy learning’ in ML. Lazy learning methods do not derive rules or relationships during learning. Rather, they compare the instances to be predicted with the knowledge base only during runtime. Learning is therefore faster and new instances can be continuously added. Eager learning is the opposite case. Models are built during learning and are usually not modifiable afterwards without repeating the learning process.

4.2.1.2. Explanatory power

Many criteria exist for judging the quality and applicability of ML models (model validation, see below). When discussing the explanatory power of models, two less formal concepts are often used: Occam's razor and the Minimum Description Length (MDL) principle.

OCCAM'S RAZOR

In 1495, the Franciscan friar and logician William of Ockham [46] put forth the following maxim in his work 'Sentences of Peter Lombard': [47] '*numquam ponenda est pluralitas sine necessitate*' ('never posit plurality without necessity', or, more freely, 'All things being equal, the simpler explanation is usually the best'). This reductionist piece of philosophy, known as 'Occam's Razor', is a useful rule of thumb when it comes to choosing between two models that perform with the same accuracy on a data set. The one that is least complex, i.e. which depends on the least attributes, will be more general and perform better on unseen data.

MINIMUM DESCRIPTION LENGTH (MDL) PRINCIPLE

In information theory, Occam's Razor is formalized to give the Minimum Description Length (MDL) principle [173]. Any given set of data can be represented as a string of symbols with length *s*. A description *d(s)* of such a string is of minimal length when it uses the least amount of symbols and this length is called Kolmogorov complexity [174] of *s*:

$$K(s) = |d(s)|$$

The MDL principle states that the best hypothesis of a data set also leads to the best compression of that data. For example, a string such as '`AAAAAAAAAAAAAAAAAAAA`' (length: 20 characters) can be represented in English as '20 times A' (length: 10 characters), which is a compression ratio of 50%.

Applied to machine learning, description of a data set is the model (or hypothesis) learned from the training set and the errors (or exceptions to the hypothesis) it makes on the entire set. During learning, the description complexity for training data decreases with the error rate (less exceptions need to be coded). Overfitted models, however, are punished because they perform much worse on test data (more exceptions need to be coded). Although the MDL principle is a sensible quality criterion, it does not allow one to find the best model. This is because the Kolmogorov complexity in a given language cannot be calculated [175].

---

[46] Nowadays, 'Ockham' is usually spelled 'Occam'.

[47] Peter Lombard (c. 1100 – July 20th 1160) was a scholastic theologian and bishop. His 'Four Books of Sentences' was the standard textbook at medieval universities.

4.2.1.3. Linear separability

Most methods classify instances by creating a space of n dimensions (where n is the number of available attributes) and defining planes that separate it into regions where a given class occurs as homogenously as possible. In the case of linearly inseparable data sets, this is impossible. The XOR problem [176] illustrates this in two dimensions. XOR is an operation on two binary variables that is only true when the two variables differ, i.e. contradict each other (Figure 24). A linear method (such as linear regression) would separate the two classes (0, 1) by trying to draw a straight line. All possible solutions (gray lines in Figure 24) fail.

in1	in2	XOR
0	0	0
0	1	1
1	0	1
1	1	0

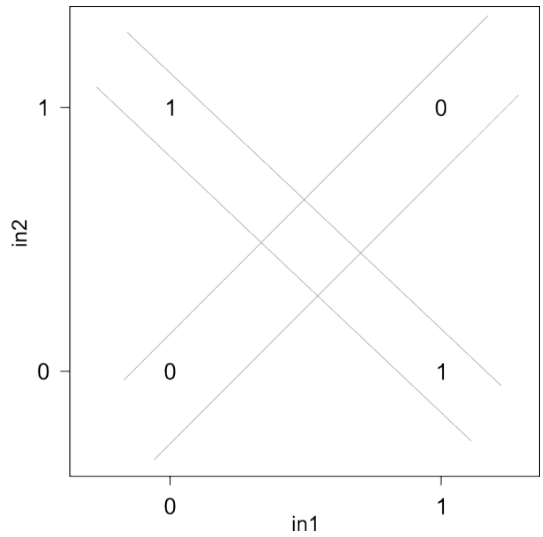


Figure 24 – XOR problem and linear separability. The XOR operation spans a problem space that cannot be separated by linear methods (decision lines, in gray).

4.2.2. Validation methods

During every modeling attempt, one is faced with two questions: is the model good enough and is there a method that performs better? In traditional statistical analysis, the quality of a model is only stated as the coefficient of determination (R^2) and it is considered valid once it crosses a certain threshold. This procedure is only applicable when a common definition of R^2 exists and its value is interpretable over different modeling techniques. This is certainly not the case in ML where numerical methods co-exist alongside classification methods.

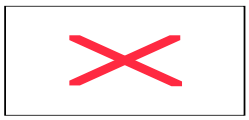
The basic principle behind all ML validation methods is using a larger part of the available data for learning (training set) and the other data to judge the performance on unseen data (test set) [177]. By modifying learning parameters and minimizing the error on training data, several models are generated and compared by their respective errors on the test set.

When comparing different learning methods to each other, part of the unseen data can be held back as a validation set, which would then be used as a test set. The best models for each method are selected

(based on performance on this validation set) and then used to predict the test set. However, this is only feasible when large data sets are available.

4.2.2.1. Skewed data and subset creation

Frequently, real-world data sets are skewed, i.e. the members of one class greatly outnumber those of the other ones. This gives rise to two problems. Firstly, if the predictive power of a method is reported as number of correct hits over the number of instances in the training set, a method can show impressive results when it constantly assigns the most frequent class to any new instance it classifies. [48] This issue is easily addressed by calculating the corrected classification rate (CCR) [49] where the predictive power of a model with n classes is reported as



where T_i is the number of correct predictions of class *i* and N_i the total number of class *i* compounds in the model. The strategy of assigning the majority class to all new instances will not perform better than a fair coin toss.

The second problem lies in the representative selection of instances for either of the sets. If a class is grossly underrepresented, it might not occur in the training set. In that case, the class will be unseen during learning and therefore not be correctly predicted. Conversely, the performance on a test set with unseen classes will be lower than if the instances were properly distributed. The work-around for this would be the stratified sampling, where one aims to proportionally represent all classes in every set instead of selecting instances at random.

4.2.2.2. Inter-rater agreement

Kappa statistics are a measure of inter-rater agreement for categorical data and considered to be a more robust measure than simple percentage calculation. They are used to compare two different assessments of the same data, i.e. when two raters (judges) assign a value to an instance, kappa statistics evaluate how well these correlate. Various methods exist, and for situations with two raters, Cohen's kappa [179] is widely used. Its value is calculated as

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

[48] This effect is especially pronounced when the most frequent class has a probability of 50% or more. It is not unlikely for real data sets to have probability peaks of 80% or more.

[49] Or any other similar measure, such as the Matthews Correlation Coefficient [178] for binary classification.

where Pr(a) is the actual agreement and Pr(e) the agreement to be expected by chance. In other words, it compares the actual success rate of the predictor (numerator) with the performance of a perfect predictor (denominator). In this respect, kappa statistics are analogous to the corrected classification rate (CCR) discussed above.

Cohen's approach can be extended to incorporate multiple raters (Fleiss' kappa [180]) by reformulating the equation to use the means of the raters' agreement:

$$\kappa = \frac{\overline{\Pr(a)} - \overline{\Pr(e)}}{1 - \overline{\Pr(e)}}$$

To calculate Fleiss' kappa for N instances indexed as i=1…N which have been assigned one of k categories indexed as j=1…k by n different raters, one first calculates the mean expected agreement per chance, Pr(e), from the proportion of all assignments to a given category j (p_j):

$$p_j = \frac{1}{Nn}\sum_{i=1}^{N} n_{ij}$$

and

$$\overline{\Pr(e)} = \sum_{j=1}^{k} p_j^2.$$

The mean of actual agreement, Pr(a), is calculated from the extent of the agreement to which raters agree on a given instance i (P_i).

$$P_i = \frac{1}{n(n-1)}\sum_{j=1}^{k} n_{ij}(n_{ij}-1)$$

and

$$\overline{\Pr(a)} = \frac{1}{N}\sum_{i=1}^{N} P_i$$

Generally, it can be said that the higher the κ value, the better is the agreement. Landis and Koch [181] give an interpretation of *κ* values which is valid for both Cohen's and Fleiss' kappa (Table 9). It is worth noting that this table is based on the personal opinion of the authors and therefore up for debate.

Furthermore, as the number of categories *k* decreases, *κ* can be shown to increase [182], showing the lack of robustness of this approach. Despite all this, both kappa statistics and Landis and Koch's table are widely used instruments in inter-rater agreement analysis.

Kappa value	Interpretation
κ ≤ 0	No agreement
0 < κ ≤ 0.2	Slight agreement
0.2 < κ ≤ 0.4	Fair agreement
0.4 < κ ≤ 0.6	Moderate agreement
0.6 < κ ≤ 0.8	Substantial agreement
0.8 < κ ≤ 1.0	Almost perfect agreement

Table 9 – Interpretation of kappa values according to Landis and Koch [181]

4.2.2.3. Diversity of sets

In analogy to assessing chemical diversity (see 4.1.6), one may assess the suitability of available data for modeling. One obvious way of doing so is to evaluate the variance of descriptors because it is pointless to look for patterns that show no variation for different classes. For some methods (e.g. decision trees, see 4.2.4), this selection is built into the algorithms. For others (e.g. neural networks, see 4.2.6) it is not.

Normalization of variables eliminates bias towards those which are on larger scales. Auto-scaling is an easily applied technique, producing descriptors scaled down to a mean $\bar{x}' = 0$ and σ' = 1:

$$x_i' = \frac{x_i - \bar{x}}{\sigma}$$

In some cases, independent variables correlate with each other such as that plotting them together gives a scatter along a straight line (Figure 25).

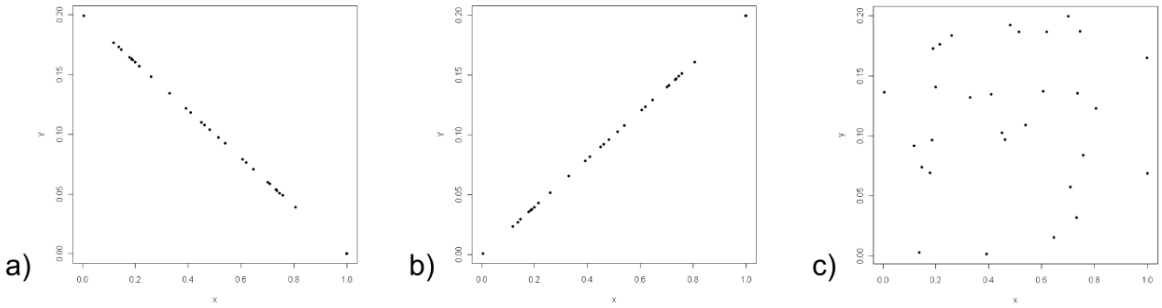


Figure 25 – Correlation of example data. a) Negative correlation (Pearson's r: -1.0), b) positive correlation (Pearson's r: +1.0), c) random data with almost no correlation (Pearson's r: -0.05)

Pearson's correlation coefficient *r* checks for negative or positive correlation (*r* approaches -1.0 or +1.0, respectively, and 0.0 for no correlation at all) between two variables *x* and *y* with means $\bar{x}$ and $\bar{y}$ for *n* instances:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]\left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right]}}$$

4.2.2.4. Cost sensitivity

Not all classes are necessarily of equal interest and two models with the same predictive power may not be equal when it comes to applicability. For instance, a virtual mass screening of compounds for leads in drug discovery may value a model that identifies more potential candidates with low precision over one that safely discards all uninteresting structures while also suppressing more possible leads. [50]

Artificially increasing samples of the highest-priced class k-fold is another way of addressing cost (cost-sensitive learning). Keeping in mind that minimizing error is at the heart of any learning scheme, it is easily seen that errors in predicting this class are also punished k-fold.

4.2.2.5. Cross-validation

Cross-validation (CV) is the most commonly practiced method of assessing the effectiveness of a model. The concept goes back to the 1950s where it was developed to evaluate the accuracy of linear regression models [183, 184]. CV aims to maximize the amount of information that can be extracted from a data set by letting instances cross over between test and training sets.

[50] This approach to model validation does not require naming an actual financial figure for cost.

K-FOLD CROSS-VALIDATION

In k-fold cross validation, a data set is partitioned into k subsets (usually k=10) [185]. Of these, k-1 are re-combined to make up a training set which is test against the last subset. This process is repeated k times until all instances have served as training and test data, thereby making sure that no classes are left out. Underrepresentation of a class in skewed data can be handled by stratified sampling.

PARTITIONING (HOLD-OUT CROSS-VALIDATION)

While technically not cross-validation because instances will not actually cross over, partitioning is often grouped with these methods and is also known as cross-validation. Here, random or stratified samples are taken from the data to form either test or training data. Test sets are usually 10-25% the size of the original data.

LEAVE-ONE-OUT CROSS-VALIDATION (LOO)

Leave-one-out cross-validation (LOO) is a special case of k-fold cross-validation [186, 187] where, in an original set of n instances, k equals n. The leave-one-out estimate is calculated by learning n models from the data, while each time removing one of the instances and testing the model against it. The ratio of failed tests is the estimate.

4.2.3. K-nearest Neighbors (kNN)

Among the plainest forms of learning is pure memorization of instances. If a 'new' instance is to be classified, it can be looked up in the table of seen instances and the once observed class is returned. Of course, as the number of available features grows and their type shifts from nominal values to continuous numerical variables, the chance of finding an exact match becomes very small.

K-nearest neighbor (kNN) algorithms address this problem by assessing the similarity of the instance to be predicted with the knowledge base [51] and the majority vote (i.e. the most frequent class) is the predicted result. The definition of neighborhood may be a fixed number of peers (Figure 26) or a geometric measure of distance where a variable number of peers are selected based on the actual distance in feature space. Usually, the number of neighbors is between five and ten. It would seem sensible to choose an odd number to prevent tied votes (and so it is often done). However, provided that the data can be robustly separated, tied votes should prove no problem, as instances would cluster and as class margins would be sufficiently far apart so that instances at the fringes would still lie in the vicinity of a homogeneous field.

[51] This follows the rationale that instances of the same class lie close to each other in feature space, much in the same way that certain neighborhoods are home to people of similar socio-economic status.

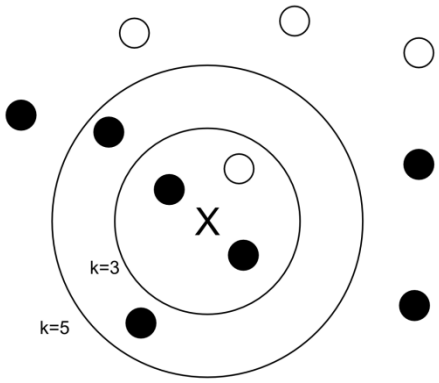


Figure 26 – Classification of an instance X using a k-nearest neighbor (kNN) approach with k=3 and k=5 nearest neighbors.

kNN algorithms are lazy learning methods: apart from storage, no processing of the data takes place. This makes training very easy and computationally inexpensive, which is an advantage compared to eager learning (such as decision tree inference (see below), where generalizations are made during training and applied later, during runtime, at low expense). A further advantage is the possibility to continually add unseen instances, thereby increasing the knowledge base and potentially performance as the model gains experience. Other models require the entire training cycle to be repeated. However, lazy learning is generally less robust and therefore more sensitive to noisy data.

4.2.4. Decision Tree Inference (DTI)

Complex problems can sometimes be tackled by recursively breaking them down into smaller sub-problems that are easily solved. This important paradigm in computer science is known as a 'divide and conquer' algorithm. In Decision Tree Inference (DTI), this type of approach is used to learn patterns, visualize data and relationships, and predict new and unknown instances.

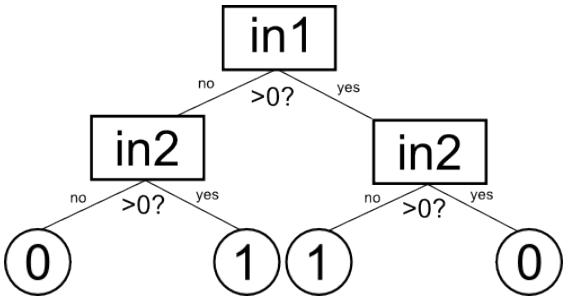


Figure 27 – Decision tree to model the XOR problem.

Trees are human readable and can therefore be used to visually convey a decision making process, quite similar to flow charts, although expert knowledge is still required to properly interpret them. Similarly, they help in uncovering patterns in new data sets. At the same time, trees are easily translated into rules by

reading the path off each leaf and translating them into programming or database query languages such as SQL (Structured Query Language, the most widely used standard). Its applicability to real-world questions is further underlined by its ability to solve linearly inseparable problems such as the XOR problem (Figure 27).

To find the optimal model for a given dataset, one would need to enumerate every possible configuration, which, considering that the complexity of the search space rises exponentially with the number of attributes and instances included, is not feasible. Instead, decision tree algorithms offer a heuristic analysis using a divide-and-conquer approach [52] in the hopes of achieving a good global model by solving simple local problems.

4.2.4.1. General description

DTI mimics the human learning and classification process by splitting a training data set into smaller and smaller subsets, with each level having higher purity than the one above. In analogy to real trees, decision trees also have nodes (or inner nodes), branches, and leaves (or leaf nodes, Figure 28). Nodes are where a given attribute is compared to a constant value and the tree branches accordingly. Nodes where no decision is made, i.e. where the path reaches a dead end, are called leaves and give a classification. In their simplest form, nodes will split two ways (binary trees), [53] although different methods exist that produce n-ary trees. [54] The depth (sometimes referred to as 'height') of a tree is the length of the path from the root node to deepest node.

A DTI algorithm will start off with the complete training set, evaluate all available attributes, and choose the one which best separates it. It then recursively proceeds to split the resulting subsets until

a) there a no more attributes,
b) the tree reaches a certain complexity, or
c) no improvements can be made by continuing to split.

When one of these break conditions is met, the node is declared a leaf that gives the class that most frequently occurs in the subset. [55]

---

[52] Divide-and-conquer algorithms are an important design pattern in computer science. Problems are recursively broken down into ever smaller problems of the same type until they become solvable.

[53] Any n-ary tree with n>2 can be transformed into a binary tree. The nodes of a complete binary tree have two or no children, a perfect binary tree has all of its leaves at the same level.

[54] A typical trinary split would be "within a range", "below the range", and "above the range".

[55] In case of equal probabilities, the class is selected at random.

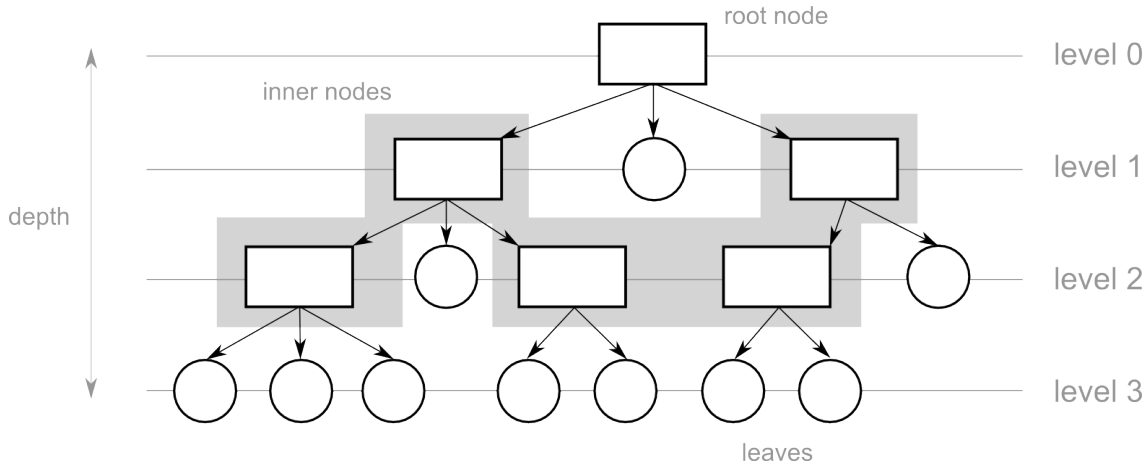


Figure 28 – Example of a decision tree with a depth of 3. Boxes are inner nodes, circles are leaves, and the root node is the only element at the top level (adapted from [188]).

4.2.4.2. Splitting Criteria

An ideal DTI algorithm creates leaves that are homogenous, i.e. they contain only instances of a single class, hence giving predictions with absolute certainty.

INFORMATION GAIN

The commonly used measure of homogeneity (or purity) was introduced in 1948 by Claude Shannon [189] as information entropy. Entropy is maximal when all classes have the same probability and minimal when a subset is made up only of members of a single class. Algorithms will therefore use entropy as an error function which they seek to minimize.

The Shannon equation is the prototypical way of calculating information entropy for a discrete random variable *X* which can take on values between x_1 and x_n (*b* is chosen arbitrarily and *0 log (0)* is taken as 0):

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i)$$

The unit of information is the bit and is not to be confused with the binary digit as it can take on fractional values. Information entropy is calculated by finding the probabilities for every class and multiplying it by their logarithms. Choice of the logarithm's base is arbitrary but is usually equal to the number of classes so the equation returns values between 0 for perfect purity and 1 for equal probabilities.

The coin toss illustrates the character of information entropy (Figure 29). In a series of tosses with a fair coin, i.e. one that is equally likely to come up with either face, the entropy amounts to 1 bit. As the coin is increasingly poised towards heads or tails, entropy converges to 0 bit.

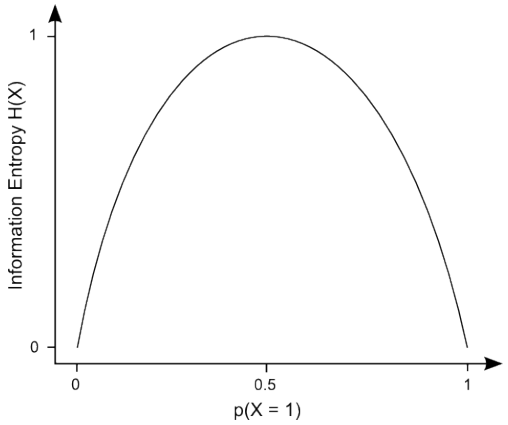


Figure 29 – Information entropy of the coin toss and the probability p(X=1) of a given face coming up.

The information entropy in the parent node is pitted against those of its children to produce the information gain. The simplest way of doing this is by subtraction.

GINI COEFFICIENT

The Gini coefficient was developed by the Italian statistician Corrado Gini [190] and is a measure of dispersion. It is most prominent in economics where it denotes the dispersion of wealth in a population, usually on a national level, and has only recently made its way into machine learning.

In decision analysis, the Gini coefficient describes the degree of slope in a Receiver Operating Characteristic (ROC) curve, a cumulative graph where the x-axis corresponds to a test or split variable and the y-axis to the percentage of instances of the same class in a data set. As the split variable progresses from its minimum to the right, more and more instances are covered until they plateau at their maximum of 100%. A variable that does not aid in discriminating between classes would run across the diagonal of the graph (line of perfect equality) and have the same discriminating power as a fair coin toss (Figure 30a). Consequently, one can gain an impression of the usefulness of the split variable by inspecting the curve and judging the slope it has. The farther it deviates to one corner, the better (Figure 30b and c). ROC curves also allow one to find the value at which no more instances of the class in question are caught, i.e. where sensitivity does not increase anymore. Setting the threshold in Figure 30b to 0.2 will include most instances and is probably preferable to a threshold of 0.4, where all instances are caught but also some of other classes.

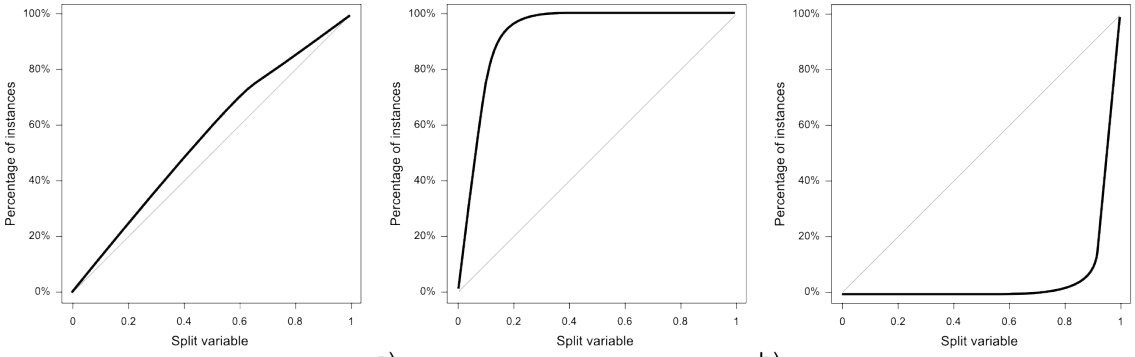


Figure 30 – Sample ROC curves

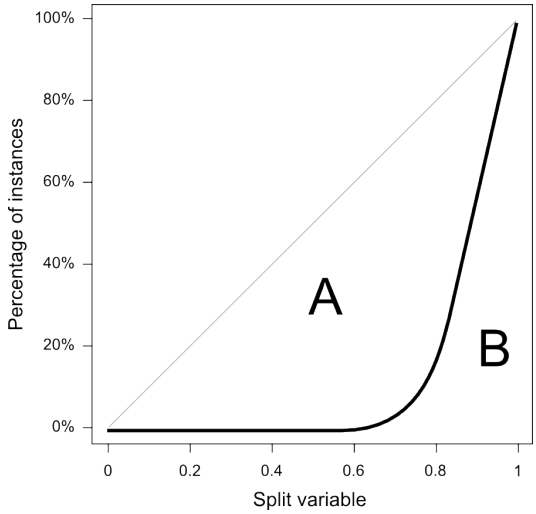


Figure 31 – Calculating the Gini coefficient. The area under the Lorenz curve (B) is divided by the area under the gray line of indifference (A+B).

The Gini coefficient formalizes the selection of the best curve. If A is the area between the line of perfect equality and the ROC curve and B is the area under the ROC curve (Figure 31), the Gini coefficient G is

$$G = \frac{A}{A + B}$$

or, taking into account that A and B cover the entire area under the line of perfect equality and therefore equals 0.5:

$$G = \frac{A}{0.5}$$

This equation works well for curves with a configuration as the one in Figure 30c but when G gives values > 1 for the mirror image Figure 30b. To remove this bias, the equation is modified to

$$G = \frac{|B - 0.5|}{0.5}$$

4.2.4.3. DTI algorithms

In order to find the optimal tree for a training set, one would have to create all possible trees and then select the best one. However, as the search space in real-world problems quickly becomes too large to do this, only heuristic algorithms exist.

ITERATIVE DICHOTOMISER 3 (ID3)

Initially described by Ross Quinlan in the 1980s in a seminal paper [191], the Iterative Dichotomiser 3 (ID3) is one of the best known DTI algorithms. ID3 is based on Occam's razor, i.e. it prefers the simpler model over the more complex (smaller trees over large ones). The algorithm builds trees by inspecting all available attributes and split points and then calculating the information gain these splits would give. The split with the highest gain is chosen and the procedure is repeated on the subsets.

C4.5 AND C5.0/SEE5

Quinlan improved ID3 to handle both continuous and discrete values, missing data, and risk/cost [192]. Furthermore, C4.5 also supports pruning (see below). While this implementation is freely available from his website (though no longer supported), Quinlan commercialized a closed-source version (C5.0 for UNIX, See5 for Microsoft ® Windows ™) that produces smaller trees at a higher speed and also added support for boosting, weighting and winnowing.

CLASSIFICATION AND REGRESSION TREES (CART)

First introduced by Breiman in 1984 [193], CART produces classification or regression trees depending on whether the dependent variable is categorical or numerical. The split criterion used is the Gini coefficient. CART produces trees by monitoring the error on the test data during growth and choosing the one with minimal error.

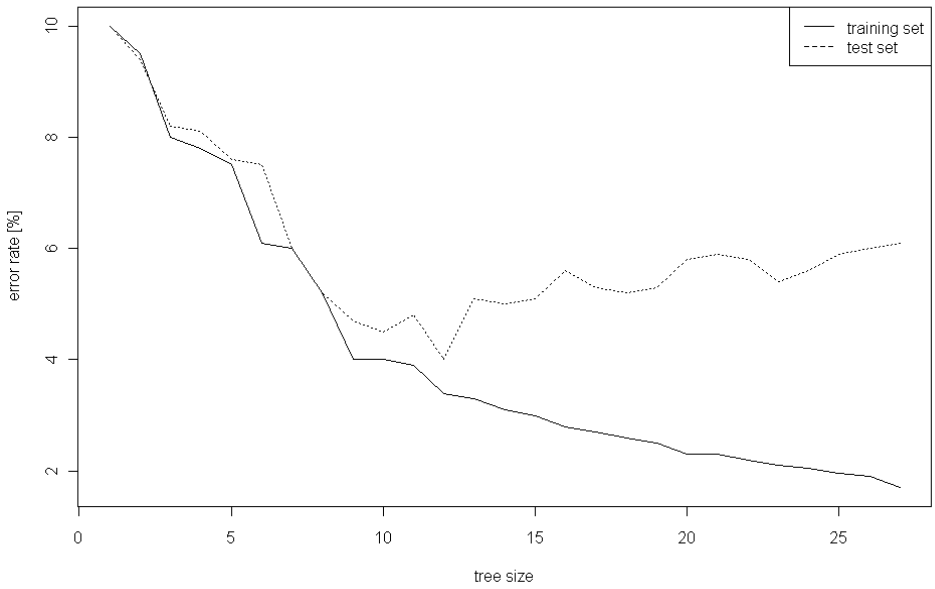


Figure 32 – Comparison of error on training set and test set (example). Training error declines steadily with tree complexity while the error on the test set reaches a minimum soon and then starts to increase again (overfitting).

CHI-SQUARE AUTOMATIC INTERACTION DETECTOR (CHAID)

The Chi-square Automatic Interaction Detector (CHAID) is the oldest of the DTI algorithms in use today and was described in 1964 [194]. Attributes are chosen with the Chi-square test and trees are subject to forward pruning.

4.2.4.4. Additional techniques in DTI

PRUNING

When presented with two models that explain the same data with the same accuracy (or error rate), one would choose the simpler one in accordance with Occam's razor. In DTI, the process of simplifying a decision tree is known as pruning. Apart from producing models that are easier to interpret, pruning also helps to avoid overfitting because the amount of information extracted from the training set is minimized. This makes the model more general and improves applicability to other (test) sets.

There are two ways of pruning a tree: forward and backward. Forward pruning is done by setting criteria that stop growth such as maximum depth, maximum number of nodes, number of instances in child nodes, and so on. Backward pruning involves predicting a test set from a maximal tree and then successively cutting branches, i.e. removing subtrees, while keeping the error rate about the same. The success of a pruning operation is measured by an improvement on the error on the test set.

A special case of backward pruning is subtree raising. The most popular subtrees, i.e. the ones with the most instances, are identified and their subtree is raised a level to replace its parent node. Any other children of the removed node are reclassified in the new tree. This procedure is resource hungry and only the most popular subtrees are examined. [56] There is no universal agreement on the usefulness of this method but it is used in some of the major DTI algorithms (C4.5, CHAID).

Pruning can also help in a post-mortem (post-hoc) attempt to decrease overfitting [195]. For this, performance on test set data is assessed and, as the tree is being pruned, constantly re-assessed in the hopes of seeing improvement on the unseen instances (at the expense of decreased performance on training data).

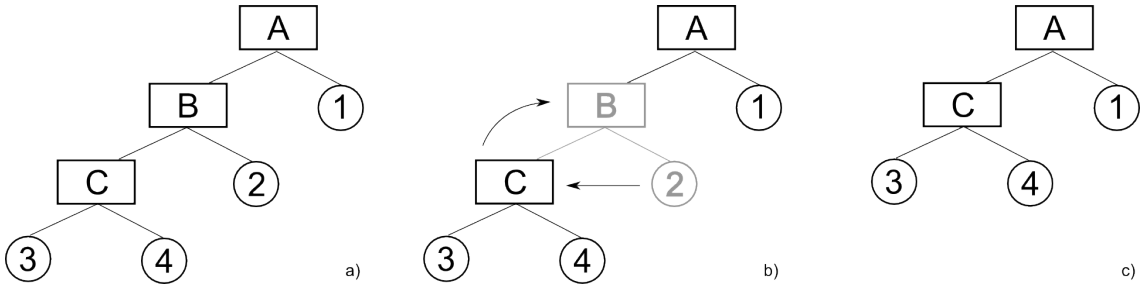


Figure 33 – Subtree raising. In the original tree (a), node B is removed (b), replaced by node C, and the instances in leaf 2 reclassified to fall into leaves 3 and 4 (c).

BOOTSTRAP AGGREGATING (BAGGING)

This procedure was proposed by Leo Breiman [196] as a method to artificially improve the accuracy [57] of regression and classification models. Given a training set of size *n*, a bagging algorithm creates *i* subsets of size *n'* < *n* by sampling the training set with replacement. All of the subsets are learned individually and the models are combined by voting (for classification models) or averaging (in regression). In linear models such as linear regression, bagging cannot be used because attributes are averaged [185, 197].

4.2.4.5. Evaluating DTI models

As with other statistical models, the quality of a tree depends on its error rate on training data and its ability to generalize, i.e. the error on test data and the tree's complexity. A DTI algorithm extracts as many features as necessary to minimize the error rate on training data. This, however, does not guarantee good performance on new data because the model was built specifically for this data set and training error rates

[56] It is important to note that subtree raising is a local modification that does not allow changes in structure to move up the tree. Considering the example in Figure 33, raising subtree C might affect the decision made on attribute A.

[57] The name refers to the analogy of pulling oneself up by one's bootstraps.

are therefore an overly optimistic estimation of future performance. [58] Furthermore, in practical settings, data are limited and test sets again make up only a fraction of the available instances. While test sets are usually a random sampling, it is likely that it is not entirely representative of the training set. Therefore, some of the predictors extracted from the data may not be present in the test set and result in higher error rates. Of course, this means that the test set contains additional information that could also be modeled if samples are chosen more adequately.

The performance and features of a DTI model must be evaluated in view of its purpose. If the tree model is intended to discover hitherto unknown patterns in data, a simpler model is certainly preferable over a complex one with lower error rates. On the other hand, if the cost associated with incorrect predictions is high, the reverse situation is preferred.

4.2.4.6. Geometrical interpretation

Decisions in DTI are – at the node level – simple comparisons of an instance's attribute to a threshold constant. Repetitions of this split the feature space into homogeneous clusters. In the case of a two-level binary tree, this corresponds to decision lines (Figure 34).

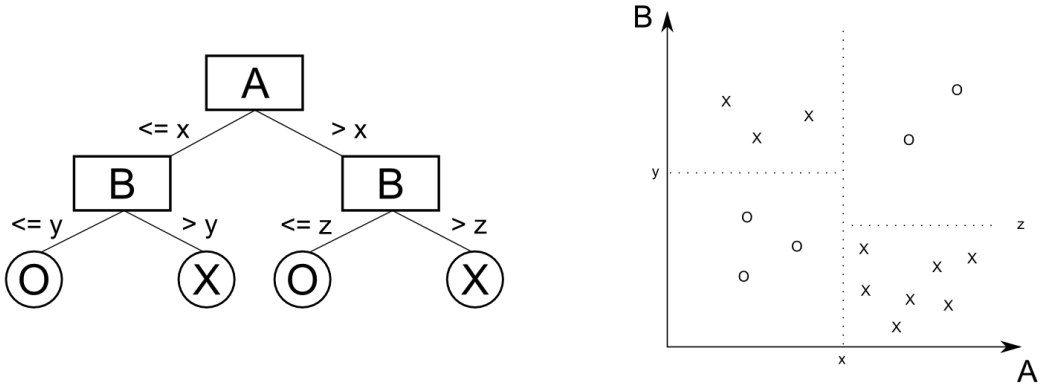


Figure 34 – Geometric interpretation of decision tree inference in a two-class problem for the classes X and O.

LEAF ERRORS

When predicting an instance with a decision tree, it is passed down the branches until it reaches a leaf. The class the leaf votes on is the most popular one during learning. In the case of a leaf that was reached during learning by 15 instances of class A and only 5 of class B, it would vote on class A. When it comes to the certainty of the prediction, one might give the overall error of the model for a given class. A more

[58] This is referred to as resubstitution error because data are re-entered into the model and predicted. The model has previous knowledge of the data set and will perform much better than on unseen instances.

precise way of doing so is to store leaf purity along with the tree structure and return this as the accuracy (that is 15/(15+5) = 75% in the previous example).

4.2.5. Random Forests

Taking the tree paradigm a step further, Breiman introduced Random Forests (RF) in 2001 [198]. It creates a pre-set number of unpruned decision trees (originally using Breiman's own CART algorithm, but others are just as valid) with a small number of different features randomly selected during learning for each tree. Classification is according to majority vote of all trees. How many trees are created in the model depends on the number of features and instances, but around ten is what is usually chosen in most studies. Of course, to avoid tied votes, an odd number is preferable. However, if class separability is high enough, tied votes occur so rarely that this constraint seems superfluous.

4.2.6. Artificial Neural Networks (ANNs)

The concept of artificial neural networks (ANN) was first proposed in 1943 by McCulloch and Pitts [199]. The idea is fundamental to bionics, the modeling of artificial systems based on biological design, and mimics the way the brain processes information. Many variations of the classic ANN model (the perceptron) exist and its applications are just as varied, ranging from pattern recognition to the control of autonomous vehicles.

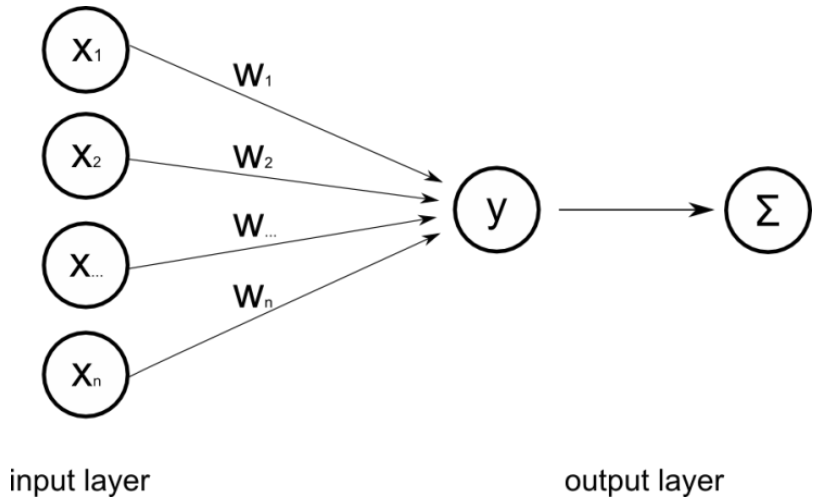


Figure 35 – Single-layer perceptron with an input layer, a weight matrix w_i, and an output layer with a transfer (*y*) and an activation (*Σ*) function.

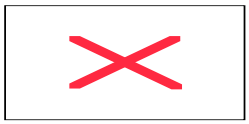
4.2.6.1. The Perceptron

The perceptron as described by Rosenblatt [200] is a direct mathematical model of the neuron and the simplest incarnation of an ANN: a feed-forward linear classifier. In analogy to the neuron, the perceptron has any number of inputs (dendrites) and only one output (axon). Information is collected in the input layer (Figure 35), modified by weights, and then directed to the output layer where the information is summed up and passed to an activation function.

Mathematically, the single layer perceptron is similar to linear regression in that it multiplies individual variables by a constant weight and then adds them up in the transfer function:

$$y = \sum_{i=1}^{n} w_i x_i$$

While linear regression stops here, the perceptron passes the value to an activation function f(y) that decides whether and to which degree the neuron fires. Figure 36a illustrates the simplest activation function (linear). Closer to biological template is sigmoid activation (Figure 36b) or the more extreme step function that models the all-or-nothing principle of the action potential in neurons. The sigmoid function in Figure 36b is given by the following equation



where Θ is the inflection point and T regulates the slope. In the step function, also known as Heaviside step function or Theta function, Θ is the threshold of activation:

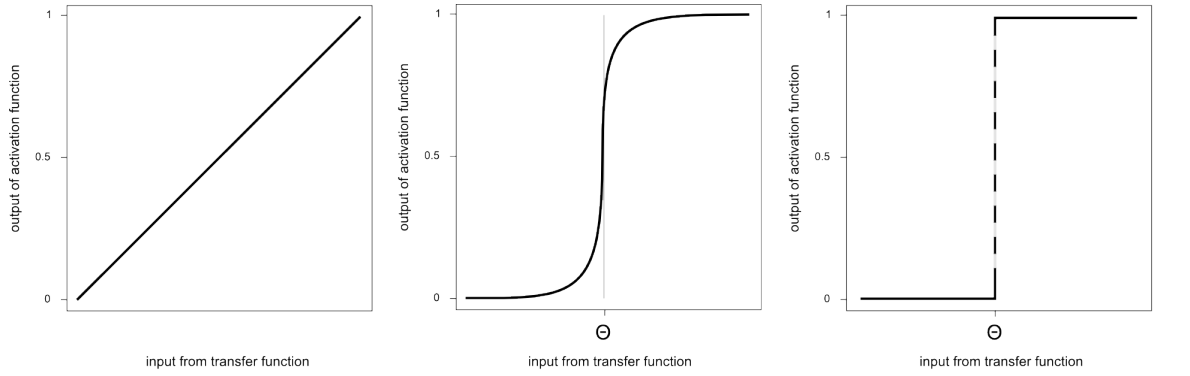


Figure 36 – Examples of activation functions: linear (a), sigmoid (b), and step function (c)

4.2.6.2. Perceptron learning

Choosing the weights w_i in a perceptron is a supervised learning process. Because of the similarities between linear regression and single-layer perceptrons, one could simply perform a least-squares analysis

of the data and substitute the calculated factors as weights. This would, in fact, yield more accurate results than the heuristic algorithms of ANN learning but would not be applicable to other network architectures.

The learning methods for ANNs are again in close analogy to the biological model. The central rule behind them is known as Hebb's Rule and states that the connection between a neuron *i* and a neuron *j* will be reinforced (that is, the weight w_{ij} is increased) when it is activated repeatedly. During learning, this weight adjustment would then follow the equation

$$\Delta w_{ij} = \eta x_i x_j$$

where η is an arbitrarily chosen learning rate. This process is repeated over the training set until a maximum number of iterations is reached or the overall error of the net reaches a minimum. A common modification of this for single-layer perceptrons is the delta rule, where η is the learning rate, t_j is the target value, v_j the actual net output, and x_i the i^{th} input:

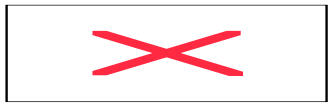
$$\Delta w_{ij} = \eta(t_j - v_j)x_i$$

Using the example topology given in Figure 35, one can formulate pseudo-code of the learning rule (Figure 37).

```
repeat for each instance i
iterations = iterations + 1
       instance_error = calculate_net(i) - expected_value(i)
       net_error = net_error + instance_error
 for each weight w
        w = (instance_error - expected_value(i)) * learning_rate
 end for
until net error converges or iterations == maximum
```

Figure 37 – Pseudo-code of a single-layer perceptron learning rule (delta rule)

It is important to avoid biasing descriptors that are on larger scales than others. Range-scaling, a modification of auto-scaling (see 4.2.2.3) with σ in the denominator replaced by the spread *r* (the difference between maximum and minimum values), yields values of a range [-0.5, 0.5]. Adding 0.5 to results shifts the range to [0, 1] and thereby all variables to comparable ranges:



4.2.6.3. The multi-layer Perceptron

Adding additional layers of neurons between the input and output layers of a single-layer perceptron, known as hidden layers (Figure 38), yields a network architecture called the multi-layer perceptron (MLP).

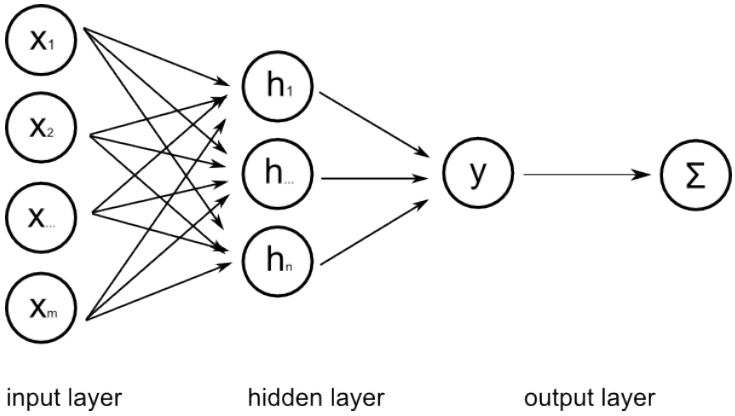


Figure 38 – Multi-layer perceptron (MLP) with one hidden layer, a weight matrix w_i, and an output layer with a transfer (y) and an activation (Σ) function.

MLPs have been proven to be universal function approximators [201], that is to say they are capable of approximating any multivariate function with any desired degree of accuracy, given that the weights and network topology are chosen properly. This is particularly noteworthy because unlike single-layer perceptrons, MLPs can model linearly inseparable data (Figure 39).

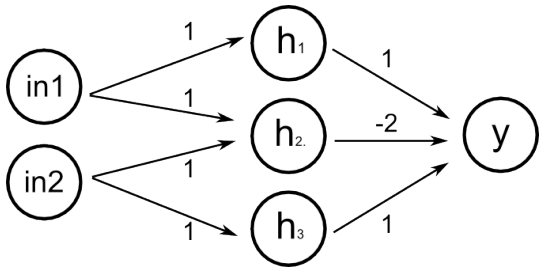


Figure 39 – XOR problem modeled in a multi-layer perceptron. The numbers along the edges are the weights assigned to the connections.

4.2.6.4. Other ANN topologies

The topologies discussed so far were feed-forward design. Recurrent variants also exist and are used in practice. Here, the output of certain neurons is fed back into the net (backwards propagation) in what is a directed cycle in graph theory. A variety of different architectures exist. One of the more common ones is the Hopfield network, where connections are symmetrical (Figure 40).

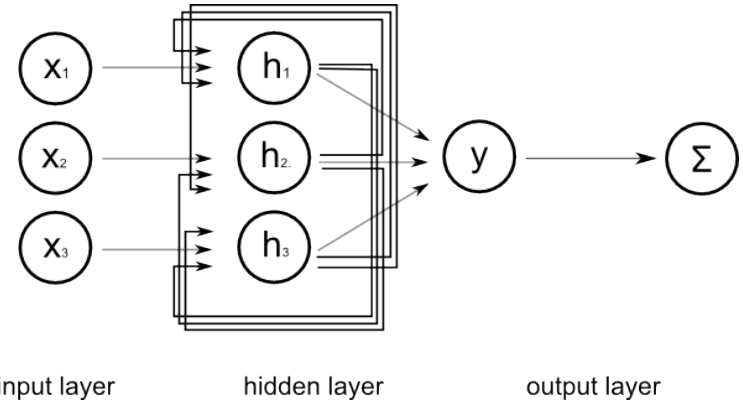


Figure 40 – Topology of a Hopfield net with 3 nodes.

4.2.6.5. Interpreting ANNs

Even though ANNs are a direct abstraction of how knowledge is represented in organisms, the rules learned cannot be read off their graphs. This is especially true of topologies with hidden layers. In single-layer perceptrons, one can judge the relative importance of inputs by comparing their weights if the training data was normalized prior to learning.

The models produced by ANNs are decision hyperplanes. In the linear case, these are planar and defined by the weights assigned to the attributes during learning. It can be stated as

$$g(x) = w^T x + w_0$$

where w^T is the vector transpose of the weights learned (determining the orientation of the hyperplane) and w_0 is a constant. Because weights are randomized in the beginning and learning is the re-orientation of the hyperplane until all classes are separated to a satisfying degree, ANNs return only one of an infinite number of solutions.

More complex architectures such as the multilayer perceptron are analogous. However, because additional layers introduce additional weights, they operate in higher dimensionalities than the original feature space. The decision boundaries therefore appear irregular (non-linear) in feature space.

4.2.7. Support Vector Machines

While ANNs are very expressive and capable of capturing non-linear relationships, computational complexity and their sensitivity to local extremes are some of their major drawbacks [29]. Support Vector Machines (SVM) remedy this situation. [59] SVMs aim to find a hyperplane through the problem space that

[59] Vladimir Vapnik presented original work on SVM learning theory in 1963 [202] and continued developing the methodology at AT&T Laboratories. This industrial setting is echoed in the orientation towards and applicability of SVMs for real-world problems [203].

not only separates all available classes but does so also with as large a margin as possible (increasing robustness).

4.2.7.1. Classification in the linear case

The geometrical solution of the simplest case (a linearly separable two-class problem) is illustrated in Figure 41. First, the outer perimeter of each class is determined, [60] similar to placing a rubber band over the instances. When released, the rubber band will snap into a position that coincides with the outer perimeter. This is a first reduction in complexity, as only the points on this polygon need to be evaluated. Next, for each edge or vertex of the outer perimeter, a plane can be found which runs parallel to the most proximal edge or vertex of the second class (the margin planes). For every pair of margin planes, a potential solution plane can be constructed which lies at an equal distance between them. These constraints limit the number of possible solution planes considerably (compared to an infinite number in ANN). The solution hyperplane finally chosen will be the one whose distance is maximal to the most proximal instances of each class.

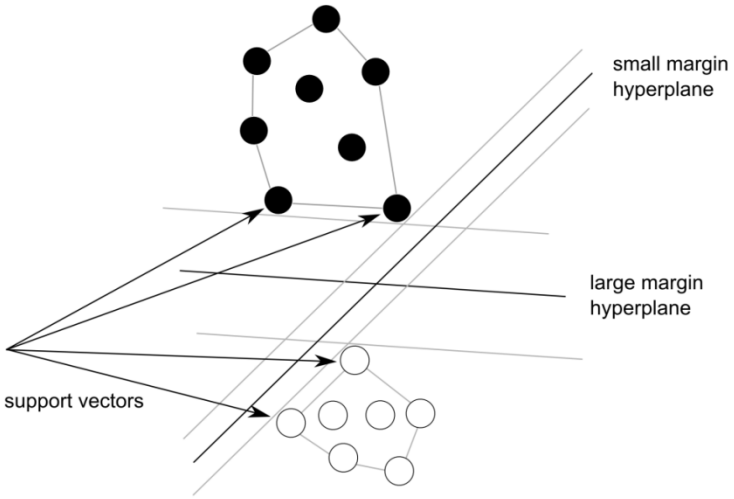


Figure 41 – Constructing a hyperplane in a two class case (example). Support vectors give the positions of the most proximal instances. Two possible solutions are given.

This geometrical solution can be formalized in computer accessible terms. Here, the algorithm is presented with training data whose instances n are points x_i in a hyper-dimensional space of magnitude m (where m is the number of attributes) with labels $c_i \in \{ -1, 1 \}$ designating class membership. The decision hyperplane is defined by a normal vector w and a translational constant b for the offset from the origin as

[60] Unless the instances are collinear (i.e. on a straight line), the outer perimeter is a convex hull, equivalent to a closed polygon. In graph theory, this is the shortest cyclical path containing all instances of the class, either as vertices or enclosed within the path. This shape also appears when all instances are maximally connected.

$$w \cdot x - b = 0$$

As was seen before, the margin hyperplanes are parallel to each other, i.e. they share the same normal vector *w*. Their equations can be written as

$$w \cdot x - b = -1; \quad w \cdot x - b = +1$$

for all points that lie on the margin hyperplanes themselves, the support vectors. All members of a class therefore satisfy one of the inequalities

$$w \cdot x - b \leq -1; \quad w \cdot x - b \geq +1$$

These can be combined using the class label as

$$c_i(w \cdot x_i - b) \geq 1$$

for all instances x_i with their assigned labels c_i. The distance of the margin hyperplanes is

$$\frac{2}{||w||}$$

The goal of an SVM learning algorithm is thus to find a vector *w*, a set of support vectors, and a constant *b* that satisfy the above mentioned conditions. This optimization problem can be stated as:

Minimize (w, b) in [ ] for any [ ] subject to $c_i(w \cdot x_i - b) \geq 1$.

If the term ||w|| is replaced with ½ $||w||^2$ in order to avoid the difficulties of working with square roots, [61] this problem can be tackled with quadratic programming, an extension of linear programming [29].

The problem can be stated in its dual form as having to find those a_i such that the expression

$$\sum_i a_i - \frac{1}{2} \sum_{i,j} a_i a_j c_i c_j (x_i \cdot x_j)$$

[61] The norm of *w* is $\|w\| = \sqrt{(x_1^2 + x_2^2 + ... + x_n^2)}$.

is maximized subject to the constraints

$$a_i \geq 0 \quad ; \quad \sum_i a_i c_i = 0$$

SVM classification can therefore model large data sets using few instances (the support vectors of the margin hyperplanes). Models are very stable and rarely subject to overfitting because only the instances at the fringe are considered and from these only very few are selected to define the margin hyperplanes, which, in turn are separated by the largest possible margins.

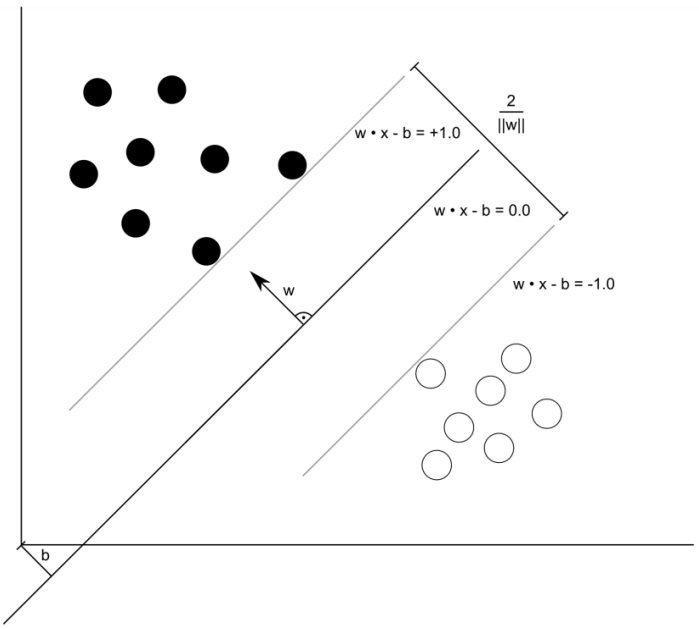


Figure 42 – Relationship of the margin hyperplanes (grey) to the decision boundary hyperplane (black) in SVM learning.

4.2.7.2. Non-linear case and the Kernel Trick

Hyperplanes are mere extensions of the decision lines encountered before. As such, they are rigid and not capable of solving tasks like the XOR problem (See Section 4.2.1.3). However, it can be shown that any problem can be solved by linear methods if only the feature space is sufficiently high-dimensional [29]. The linear decision boundaries in such high-dimensional spaces appear crooked or wiggly when transformed back. Unfortunately, as the number of dimensions *d* approaches the number of instances *n*, not only does computational effort increase to an unmanageable magnitude, but also the danger of overfitting.

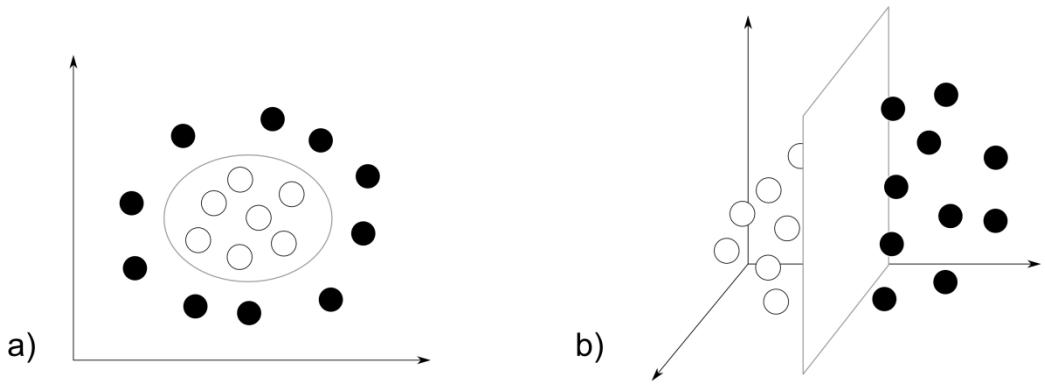


Figure 43 – Graphical representation of the kernel trick. Classification tasks that are non-linear in low dimensionality (a) may become solvable for linear algorithms when transformed to higher dimensions (b).

For SVMs, the situation is different. As was seen before, models are seldom overfitted because only a very few instances are needed to represent class boundaries and margins are maximized. Decision boundaries are therefore not very flexible. The computational cost is also not a problem. The dot product in the equations discussed in Section 4.2.7.1 can be replaced by other (kernel) functions that are far less tedious to compute. In fact, the equations can be solved before the transformation to higher dimensionality takes place. This is known as the Kernel Trick. [62] An overview of commonly used kernel functions is given in Table 10.

Kernel function	Description
$k(x,x') = (x \cdot x')^d$	homogeneous polynomial
$k(x,x') = (x \cdot x' + 1)^d$	heterogeneous polynomial
$k(x,x') = \exp(-\gamma \|x - x'\|^2)$	radial basis function (RBF)

Table 10 – Commonly used kernel functions.

4.2.7.3. Parameter optimization

Several meta-parameters define the learning properties of SVMs and their proper settings have great influence on the predictive accuracy of the final model. Unfortunately, there are no practical standard settings or guidelines, and only experimentation can hint at good settings. A common approach to estimate

[62] A year after Vapnik's paper on SVM learning, Aizerman proposed kernel transformations [204]. It was not until the 1990s that the two approaches were successfully combined to tackle real-world problems with SVMs.

these parameters is a grid search over a $\log_2$ space with uniform resolution. An initial screening with the commonly used kernel function RBF (Table 10) can be performed by taking 20 samples for C and γ (a parameter of the kernel function) so that $log_2C = \{-5, \ldots, 15\}$ and $log_2\gamma = \{-15, \ldots 3\}$. Contour plots of C, γ, and a measure of success (such as the CCR, see 4.2.2.1) can help identify maxima which may then be magnified by contracting the range around the regions of interest and increasing resolution. Section 7.1.3 of the appendix gives the listing of an application to perform grid searches.

SVM learning itself is time-consuming and even more so when performed repeatedly, especially using k-fold cross-validation. If computational expense during learning is an issue, one can consider a pattern searching approach. Here, a combination of parameter values is chosen randomly and searches are performed with hill-climbing algorithms that continue to move in a direction with upward slope (i.e. higher success rate). Even with backtracking, hill-climbing is prone to be stuck in local maxima, confusing the next best hill with the big mountain possibly just a valley away.

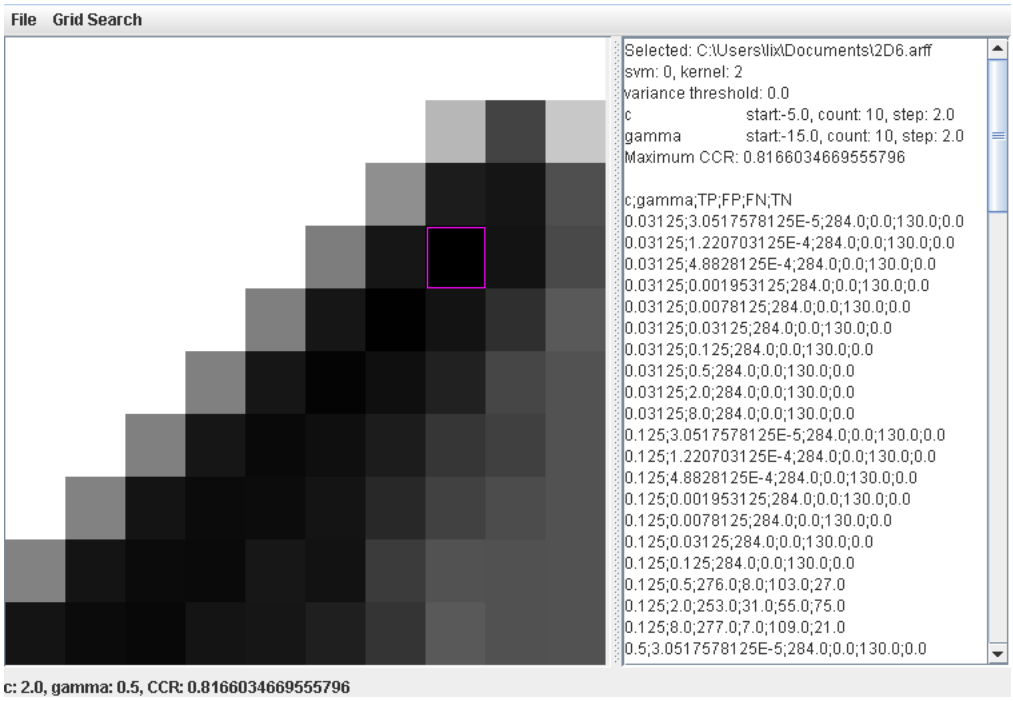


Figure 44 – Coarse grid search in a $\log_2$ space with uniform resolution for the learning parameter cost (C, vertical axis) and the radial basis function parameter γ (gamma, horizontal axis). Combinations are evaluated by their corrected classification rate (CCR) and the most successful pair is highlighted.

4.2.8. Feature Selection

It is immediately apparent that models can be learned best with a set of features relevant to the problem, e.g. when a human expert selects them manually, based on knowledge and understanding of the domain. While most ML methods assign some form of weight to attributes during training (kNN being an exception),

others completely discard features as they are being trained (e.g. DTI). Often it is not possible, even for a human expert, to identify the important characteristics of a problem, and the subsequently overly large set of instances causes several problems.

Firstly, some algorithms are more sensitive to noise than others (e.g. kNN vs. SVM). Here, irrelevant attributes cause performance to deteriorate significantly. [63] Secondly, spurious features unnecessarily complicate a problem's feature space, leading to a potentially catastrophic increase in computational complexity. The phenomenon of problem space growing as an exponential function of dimensionality is known, somewhat informally, as the Curse of Dimensionality (coined by Richard Bellman in 1961). Not only does this give a computational penalty. The danger of overfitting also increases as sample size is outnumbered by dimensionality. For these reasons, ML is often preceded by feature selection – the process of reducing the number of attributes. Two different approaches exist: filter methods vs. wrapper methods.

Wrapper methods employ an ML algorithm (possibly even the same as is to be used during learning itself) to select a subset of features. For example, DTI can be used with the entire feature set, removing the attribute in the root node and repeating the process until no attributes are left. This yields a ranked list ordered by performance. A ranking can also be made according to the magnitude of weights learned in numerical methods such as ANNs on normalized data. Filter methods, on the other hand, examine the general characteristics of the data. Numerical (and normalized) data can be ranked by variance and subsets selected via a threshold (0.05 is a common measure). Correlation coefficients (see 4.2.2.3) may be used to select features with high correlation with the class but low inter-correlation.

Regardless of the method chosen, subsets can be built in basically two different ways: either by starting out with an empty set and adding to it (forward selection) or with the entire set and removing from it (backward elimination) until a specified stopping criterion is met (subset size, inter-correlation, etc.). This can be visualized as a graph traversal (Figure 45). Generally, backward elimination yields slightly better results while forward selection produces smaller subsets [177].

Of course, the optimal balance between subset quality and size can be found by evaluating every node. However, the Curse of Complexity being what it is, calculating this quickly becomes impractical (even merely enumerating the graph can be too expensive, let alone storing the visited nodes for backtracking). Luckily, the field of computer science has many sophisticated algorithms to offer. Beam search, where the

[63] Studies with random variables (maximal irrelevance) introduced in standard datasets have shown a decrease by 5-10 % in C4.5 [177]. Note that this is an algorithm with 'built-in' feature selection. At a sufficiently deep node of the tree, only a small set of instances will be left to build a decision on and even random variables can start to appear superior to 'relevant' attributes.

subset size is limited, or best-first search, where only ‘promising’ nodes (as evaluated by a heuristic function) [64] are expanded, are just two examples.

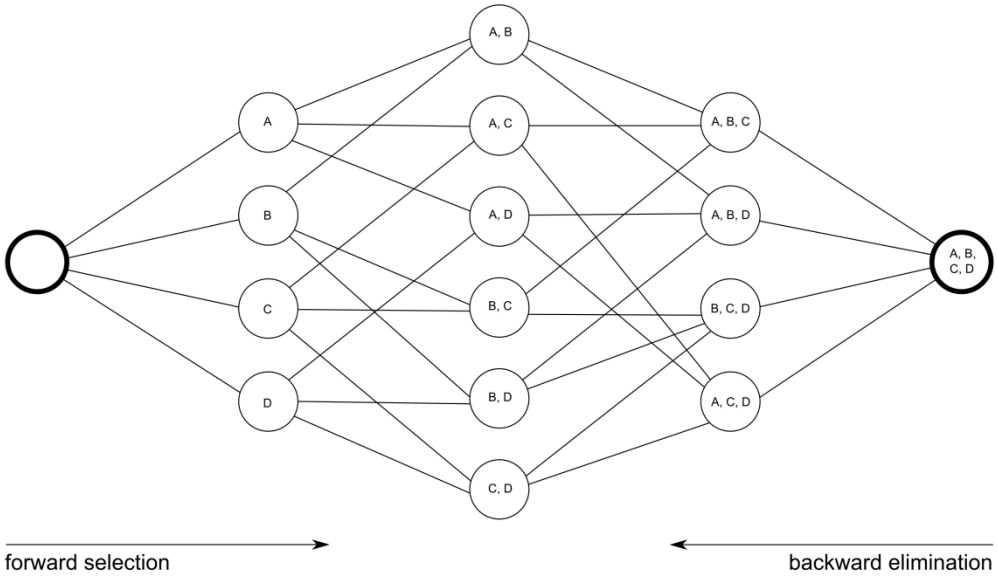


Figure 45 – Forward selection, starting with an empty set, and backward elimination, starting with the complete set, for four attributes (A, B, C, and D) represented as a bi-directional graph.

[64] In search algorithms, heuristic functions rank alternatives based on information available at a given time. Heuristic functions that are admissible (i.e. ones that never overestimate) can be shown to always find optimal solutions, given enough time. These form the class of A* algorithms, of which best-first search is an example [29, 188].

5. Projects

5.1. Prediction of drug transport and metabolism

5.1.1. Development of decision tree models for substrates, inhibitors, and inducers of P-glycoprotein

Felix Hammann [1], Heike Gutmann [1], Ursula Jecklin [1], Andreas Maunz [2], Christoph Helma [2], Juergen Drewe [1*].

[1]: Department of Clinical Pharmacology and Toxicology, University Hospital Basel, University of Basel, Switzerland

[2]: Freiburg Center for Data Analysis and Modelling, Albert-Ludwigs-University Freiburg, Germany

.

Corresponding Author (*):
Prof. Juergen Drewe
Department of Clinical Pharmacology and Toxicology
University Hospital of Basel
Petersgraben 4
CH-4031 Basel Switzerland
Email: juergen.drewe@unibas.ch
Phone: +41-61-265 3848
Fax: +41-61-265 8581

Current Drug Metabolism 2009;10(4);339-346

Abstract

In silico classification of new compounds for certain properties is a useful tool to guide further experiments or compound selection. Interaction of new compounds with the efflux pump P-glycoprotein (P-gp) is an important drug property determining tissue distribution and the potential for drug-drug interactions. We present three datasets on substrate, inhibitor, and inducer activities for P-gp (n = 471) obtained from a literature search which we compared to an existing evaluation of the Prestwick Chemical Library with the calcein-AM assay (retrieved from PubMed). Additionally, we present decision tree models of these activities with predictive accuracies of 77.7 % (substrates), 86.9 % (inhibitors), and 90.3 % (inducers) using three commonly used algorithms (CHAID, CART, and C4.5). We also present decision tree models of the calcein-AM assay (79.9 %). Apart from a comprehensive dataset of P-gp interacting compounds, our study provides evidence of the efficacy of logD descriptors and of two algorithms not commonly used in pharmacological QSAR studies (CART and CHAID).

KEYWORDS

P-glycoprotein, MDR1, Multidrug resistance, Calcein AM assay, QSAR, decision trees

CONFLICT OF INTEREST

Financial support for this project was partially provided by the EU FP7 project "OpenTox" (Contract Number Health-F5-2008-200787).

Introduction

See Section 3.3.1.3.

Materials and Methods

Quantitative Structure Activity Relationships (QSAR)

See Section 3.1.2.3.

Data set

We based our dataset on a mass screening based on the calcein-acetoxymethyl (calcein-AM) assay of the Prestwick Chemical Library retrieved from PubMed [205]. Structures were cross-checked in a search of PubMed, yielding a total of 471 compounds. For these structures, we give activities in the three classes 'inducer', 'inhibitor', and 'substrate', depending on whether publications exist that document activity ('ACTIVE') or inactivity ('INACTIVE') (Table 11). When the search was inconclusive, compounds were tagged with 'UNKNOWN'. Activity in this context means having a given property, regardless of the underlying molecular mechanism. The datasets we used in numerical analysis consist of the subsets of compounds with known activity or inactivity (i.e. instances labeled 'UNKNOWN' were excluded) and complete set of descriptors (i.e. no missing values were allowed).

	Active	Inactive	Total
Inducer	49	21	70
Inhibitor	217	123	340
Substrate	163	95	258
Total	429	239	668

Table 11 – Results of a literature search of compounds with P-glycoprotein activity

Descriptors

See Section 4.1.7.

Assessment of chemical diversity

See Section 4.1.6.

Calcein-AM assay

We chose data generated with the calcein-AM assay as *in vitro* control data for our models that was developed to assess the interaction of a compound with P-gp *in vitro* [205]. Calcein is a fluorescent dye, which enters living cells when substituted with an acetomethoxy group (calcein-AM) by diffusion. The functional group is then removed by intracellular esterases, allowing free calcein to chelate with calcium. The highly negatively charged green fluorescent calcein is rather polar, preventing it to leave cells by simple diffusion. P-gp is able to extrude calcein [206], making calcein expulsion a widely used measure of P-gp mediated multidrug resistance [207].

The results of a mass screening based on the calcein-AM assay and the Prestwick Chemical Library are available from PubChem. The data set consists of 778 molecular structures in SMILES format, along with changes in fluorescence in the assay, and preliminary classification into active (n=353), inactive (n=239), and inconclusive (n=186) compounds based on pre-determined thresholds [208].

Decision Tree Inference

See Section 4.2.4.

Data preparation

Prior to analysis, we prepared tables for each endpoint (substrate, inducer, and inhibitor activities extracted from the literature, as well as activity in the calcein-AM assay) along with descriptor values computed previously. Compounds with an incomplete array of descriptors were excluded from analysis, that is, no missing values were accepted.

Model validation

Throughout this paper, 10-fold cross validation was used. See Section 4.2.2.5.

Corrected Classification Rate (CCR)

See Section 4.2.2.1.

Kappa statistic

See Section 4.2.2.2.

Software

ChemAxon Marvin (Marvin 5.0.4, 2008, http://www.chemaxon.com) was used for characterizing chemical structures, substructures, and ChemAxon Calculator Plugins were used for structure property calculation. Additional descriptors were calculated with the open-source cheminformatics package Chemical Development Kit [209] (version 1.0.4, 2008, http://sourceforge.net/projects/cdk).

Decision trees grown with the CHAID and CART algorithms were done in SPSS 15.0 for Microsoft Windows. For the C4.5 algorithm, the original software (release 8, http://www.rulequest.com/Personal) was used.

Results and Discussion

Chemical diversity

We calculated the Tanimoto coefficient for the result of our literature search and arrived at an overall value of 0.34. This is an indication that these sets are reasonably diverse.

Comparison of Calcein-AM assay and literature

Of the compounds given in the original assay, we found 101 in literature. A cross-tabulation analysis revealed an agreement of 75 %, which is moderate in analysis with kappa statistics (κ: 0.043). This shows the calcein-AM assay to be prone to false negative predictions. Due to a large spread between active and inactive compounds, changing the thresholds to broaden the number of inclusive outcomes (thereby decreasing the number of active and inactive components) does not improve performance.

Performance of Decision Tree Models

A summary of CCRs for the decision tree models created is given in Table 12.

	CHAID	CART	C4.5
Substrate	**77.7 %**	77.0 %	70.3 %
Inhibitor	83.0 %	**86.9 %**	67.3 %
Inducer	**90.3 %**	85.0 %	63.0 %
Calcein-AM	73.3 %	**79.9 %**	60.8 %

Table 12 – Corrected Classification Rate (CCR) of decision tree models (performance on the test datasets generated during cross-validation) built with three different methods (maximum depth: 10, minimum parent node size: 10, minimum child node size: 5) for a literature search of compounds with P-glycoprotein activity along with models for a mass screening using the calcein-AM assay. Models with highest performance are shown in boldface.

Calcein-AM assay

A tree grown with the CART algorithm (Figure 46) had a CCR of 79.9 % with a depth of 6, 12 inner nodes, and 17 leaves. Discerning descriptors were the polar surface area, the number of hetero and aromatic rings, the resonant count, partitioning coefficients at pH 3.0 and 10.0, the smallest ring size, as well as the connectivity indices (Wiener, Randic, Kier kappa shape).

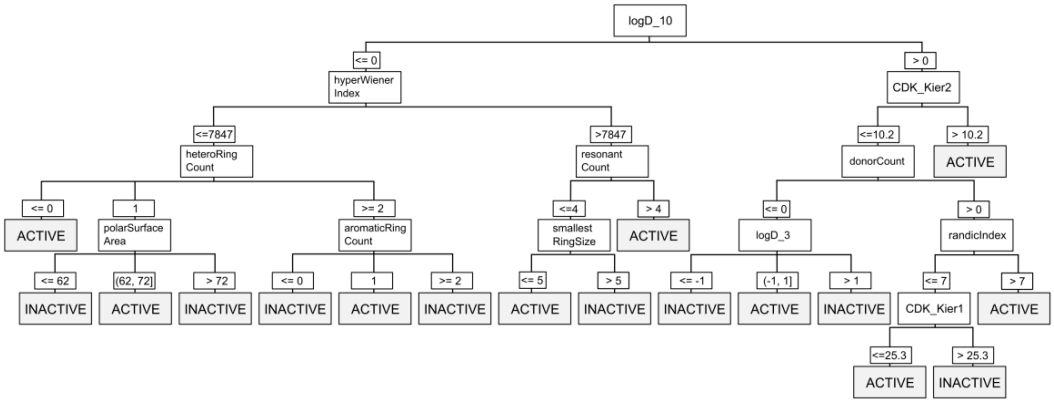


Figure 46 – Decision tree model of 434 compounds with P-glycoprotein activity determined using the calcein-AM assay. The corrected classification rate amounts to 79.9 %.

Results from the literature search

During our screening of literature, we have chosen to differentiate between the three classes of substrates, inhibitors, and inducers. Membership of these classes, however, is not exclusive, i.e. a compound may belong to more than one class (consider verapamil, which is known to be both an inhibitor [210] as well as a substrate [211] of P-gp) Our goal was to provide separate models with higher predictivity so that they may find application in different areas of drug discovery. Combining endpoints would have introduced unnecessary bias.

SUBSTRATES

With a CCR of 77.7 %, the tree grown with the CHAID algorithm (Figure 47), depth: 5 levels, inner nodes: 7, leaves: 9) performed best and included the first and third Kier kappa shape descriptor, Wiener polarizability, the number of hydrogen acceptors, the number of aromatic and fused aliphatic rings, and bond count.

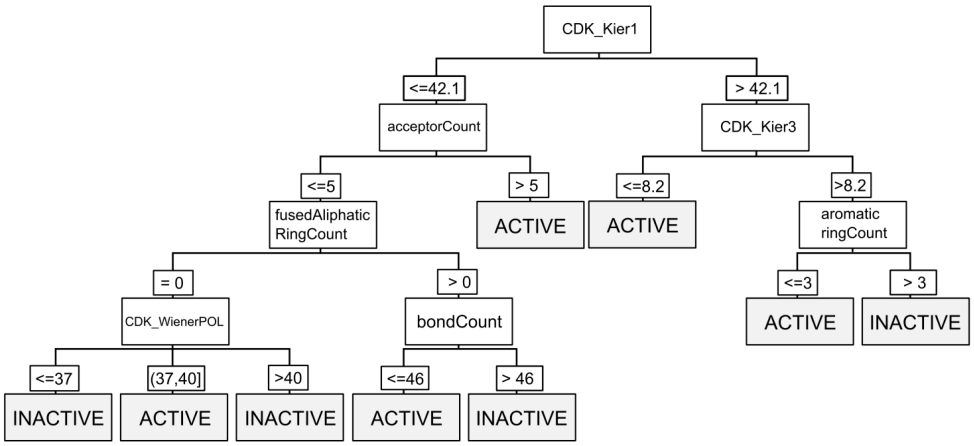


Figure 47 – Decision tree model of 229 P-glycoprotein substrates with a corrected classification rate of 77.7 %.

INHIBITORS

The CART algorithm performed best with a CCR of 86.9 %, yielding a tree with a depth of 8, 16 inner nodes, and 17 leaves (Figure 48). Discriminant descriptors include the partitioning coefficients (pH 2.0, 8.0, 10.0, and 13.0), XlogP, the counts of aromatic bonds, total bonds, ring atoms, ring bonds, and carbo and hetero rings, pi energy, Dreiding energy, and the third Kier kappa shape index.

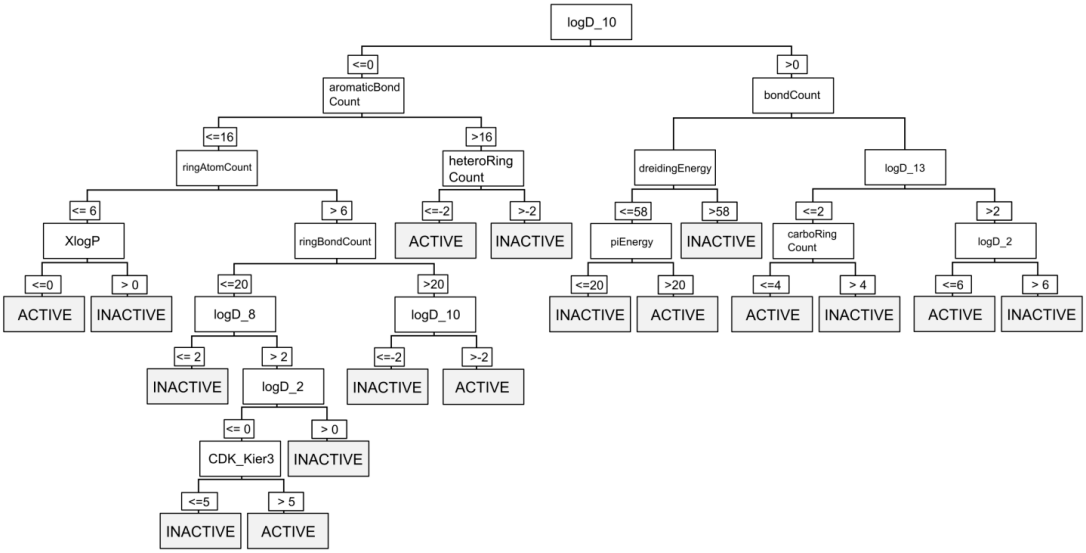


Figure 48 – Decision tree model of 298 P-glycoprotein inhibitors with a corrected classification rate of 86.9 %.

INDUCERS

A tree grown with the CHAID algorithm (Figure 49) and a CCR of 90.3 % showed the best results (depth of 3, with 3 inner nodes, and 6 leaves). The third Kier kappa shape index, ring count, and the partition coefficient at pH 4.0 appeared in the model.

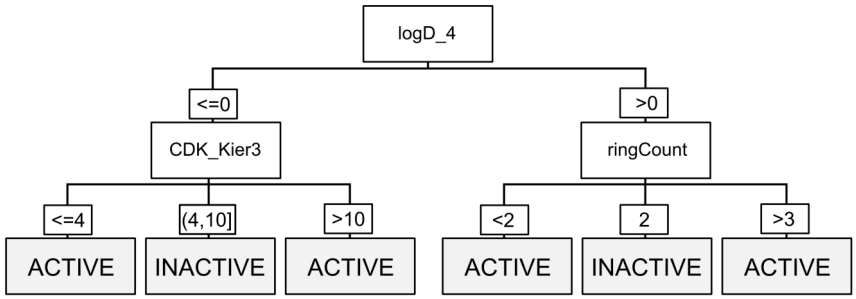


Figure 49 – Decision tree model of 61 P-glycoprotein inducers with a corrected classification rate of 90.3 %.

Discussion

In this work, we present a structurally diverse set of compounds with known P-gp interaction status (inhibitors, inducers, substrates). We show that DTI models perform well in the prediction of activities, with CCRs ≥ 0.7 for the best models. They rely on different sets of easily accessible descriptors, and are potentially useful in early drug discovery and the identification of structure-activity relationships. As has been pointed out before, combinatorial approaches to P-gp modeling currently appear superior to pharmacophore-based discovery, especially as the latter is used mostly after chemical synthesis and *in vitro* assays have already been performed [212, 213].

Evaluation of the calcein-AM assay

With 75% agreement with our literature search, we consider the calcein-AM assay only a moderately useful estimator of P-gp activity. This is most likely because of its inability to distinguish between substrates and inhibitors. We also suspect the discrepancy to arise from a greater accuracy of literature searches compared to mass screenings. There is debate over the validity of literature searches and comparability of results. For instance, Polli and colleagues [214] have shown differences in classification between *in vitro* assays of P-gp. We also feel that no single *in vitro* assay properly assesses P-gp activity. Pooling results should remedy this by allowing a wide variety of assays to enter the picture and allowing for the proper assay to be used for given compounds. The DTI model we present here is accurate enough (CCR = 79.9 %) to find application as a virtual pre-screening tool that reduces the number of test candidates in a mass screening.

Model interpretations

In our models for P-gp substrates, low bond counts, a small number of aromatic rings, many hydrogen bond acceptors, and low connectivity predominate. This suggests that small molecules with a relatively high capacity for polar interactions are good substrates. This finding is in accordance with other studies of P-gp substrates [215, 216]

The complex biology of P-gp inhibition (direct, via disturbance of ATP binding or hydrolysis, or through perturbance of the cellular membrane resulting in poor drug-transporter binding and accessibility [217]) is echoed in our second model. Descriptors of lipophilicity predominate, as do indicators of molecular rigidity (DREIDING energy, energy of π orbitals). Lipophilicity as a rate-limiting step in P-gp interaction has long been proposed by Seelig and Landwojtowicz [218] and is of great importance for mechanisms of interaction as outlined above.

The model for P-gp inducers is much less complex, since only a few compounds were available and, as a consequence, fewer structural requirements could be extracted. P-gp substrates can act as inducers, and structures of inducers are hence equally diverse as for substrates, ranging for St. John's Wort [219] and other psychoactive drugs [220] to HIV protease inhibitors [221]. This cross-selectivity leads to the appearance of similar descriptors in both models (Kier connectivity, logD, and ring count).

In the literature, concerns were raised as to the stability of DTI and the need for careful sampling of instances for training and test sets [222]. However, these concerns only apply in situations where small perturbations in the selection of instances greatly influences the models learned. During our work on the data presented here, we found no indication for this and used random sampling during cross-validation.

Conclusions

Machine learning techniques have become popular in QSAR studies of P-gp interaction. Support vector machines, k-nearest neighbor models, Bayesian and neural networks have shown promising results, sometimes outperforming DTI [212, 223, 224]. Poorer results in DTI could stem from the wide-spread use of the C4.5 algorithm, which consistently trails behind CART and CHAID in this study (Table 12). As with all machine learning techniques, no single one is right for every job, and this holds true for DTI algorithms as well. The application of other modeling techniques to our dataset would make an interesting subject of further studies.

The importance of lipophilicity is apparent in our models. Strikingly, however, traditional logP (XlogP) appears only in the inhibition model (in the second to last level), and logD, the measure of lipophilicity in ionizing environments, is featured prominently in all models. While it is debatable whether ionization is important for P-gp active compounds once they contact their target, it is clear that protein binding, membrane passage, and solubility and stability in compartments with differing pH affect a compound's

ability to reach the target in the first place. Our results indicate that logD is a highly suitable descriptor which could be attractive for QSAR studies of other biological systems.

5.1.2. Classification of Cytochrome P_{450} activities using machine learning methods

Felix Hammann [1], Heike Gutmann [1§], Ulli Baumann [1], Andreas Maunz [2], Christoph Helma [2], Juergen Drewe [1*].

[1]: Department of Gastroenterology & Hepatology, University Hospital Basel, University of Basel, Switzerland

[2]: Freiburg Center for Data Analysis and Modelling, Albert-Ludwigs-University Freiburg, Germany

[§]: present address: Novartis Institutes for Biomedical Research, Drug Metabolism and Pharmacokinetics, Basel, Switzerland

.

Corresponding Author (*):
Prof. Dr. Juergen Drewe
Department of Gastroenterology & Hepatology
University Hospital of Basel
Petersgraben 4
CH-4031 Basel Switzerland
Email: juergen.drewe@unibas.ch
Phone: +41-61-265 3848
Fax: +41-61-265 8581

Molecular Pharmaceutics 2009;6(6);1920-1926

Abstract

The cytochrome P_{450} (CYP) system plays an integral part in the metabolism of drugs and other xenobiotics. Knowledge of the structural features required for interaction with any of the different isoforms of the CYP system is therefore immensely valuable in early drug discovery. In this paper, we focus on three major isoforms (CYP 1A2, CYP 2D6, and CYP 3A4) and present a dataset of 335 structurally diverse drug compounds classified for their interaction (as substrate, inhibitor, or any interaction) with these isoforms. We also present machine learning models using a variety of commonly used methods (k-nearest neighbors, decision tree induction using the CHAID and CRT algorithms, random forests, artificial neural networks, and support vector machines using the radial basis function (RBF) and homogeneous polynomials as kernel functions). We discuss the physico-chemical features relevant for each endpoint and compare it to similar studies. Many of these models perform exceptionally well, even with 10-fold cross-validation, yielding corrected classification rates of 81.7 to 91.9 % for CYP 1A2, 89.2 to 92.9 % for CYP 2D6, and 87.4 to 89.9 % for CYP3A4. Our models help in understanding the structural requirements for CYP interactions and can serve as sensitive tools in virtual screenings and lead optimization for toxicological profiles in drug discovery.

Introduction

The Cytochrome P_{450} system

See Section 3.2.

Quantitative Structure Activity Relationships (QSAR)

See Section 3.1.2.3.

Materials and Methods

Dataset

Our dataset is based on a list of FDA approved small molecule drugs (n = 1436, date of access: 1st of January 2008) from the University of Alberta's DrugBank database [225]. We used the DRUGDEX ® system (Thomson Reuters, http://www.micromedex.com, date of last access: 1st of May 2008) to check for interactions (as substrate, inducer, and/or inhibitor) with CYP 1A2, 2D6, and 3A4. A summary is given in Table 13. Compounds where interaction was documented were labeled 'ACTIVE' whereas all other ones where labeled 'INACTIVE' when the interaction status was either unknown or known to be inactive. Compounds with at least one activity or documented absence of activity were included in the final data set.

CYP	substrates	inhibitors	inducers	total active
1A2	66 (18.7 %)	33 (9.3 %)	7 (2.0 %)	88 (24.9 %)
2D6	99 (28.0 %)	77 (21.8 %)	0 (0 %)	130 (36.8 %)
3A4	242 (68.6 %)	78 (22.1 %)	27 (7.6 %)	264 (74.8 %)

Table 13 – Number of compounds (percent of total (n=353)) with given activity for each of the cytochrome P450 (CYP) isoforms studied. The final column contains the number of compounds with any of these activities.

Assessment of chemical diversity

See Section 4.1.6.

Machine Learning (ML) Methods

k nearest neighbor (kNN) algorithm

See Section 4.2.3.

Decision Tree Induction (DTI)

See Section 4.2.4.

Random Forests (RF)

See Section 4.2.5.

Artificial Neural Networks (ANN)

See Section 4.2.6.

Support Vector Machines (SVM)

See Section 4.2.7.

Corrected Classification Rate (CCR)

See Section 4.2.2.1.

Model Validation

See Section 4.2.2.5.

Descriptors

See Section 4.1.7.

Data Preparation

We calculated a total of 118 descriptors (Table 14) based on the structures deposited at the DrugBank server. Compounds for which some descriptors could not be computed were removed (i.e. no missing values were allowed). Some of the algorithms are susceptible to over-estimating the effects of features which are on a higher numerical scale than others (e.g. number of hydrogen bond acceptors vs. molecular mass) [226]. To avoid this bias, we normalized the entire data set to a range of [0, 1]. Furthermore, feature selection was performed for all algorithms except DTI and RF (which implicitly do so themselves), since they produce better results when extraneous or irrelevant information is removed *a priori* [227]. We therefore reduced the feature space to a set of 19 features with the least correlation between them using best first forward search (a greedy hill-climbing algorithm) with backtracking [177]. Lastly, classes were re-coded to +1.0 for active compounds and -1.0 for inactive compounds in SVM models and 0.0 for inactive compounds in ANN models, respectively.

Class	ChemAxon	Chemistry Development Kit
charge analysis	Hydrogen bond acceptor and donor counts and sites (acceptorCounts, acceptorSiteCount, donorCount, donorSiteCount), DREIDING energy, partitioning coefficients for pH 0-14 (logD_0, logD_1, logD_2, logD_3, logD_4, logD_5, logD_6, logD_7, logD_74, logD_8, logD_9, logD_10, logD_11, logD_12, logD_13, logD_14), molecular polarizability (molPol), topological surface are (topologicalSurfaceArea), van-der-Waals surface area (vdwsa)	Charged Partial Surface Descriptors (PPSA-1, PPSA-2, PPSA-3, PNSA-1, PNSA-2, PNSA-3, DPSA-1, DPSA-2, DPSA-3, FPSA-1, FNSA-1, FPSA-2, FNSA-2, FPSA-3, FNSA-3, WPSA-1, WNSA-1, WPSA-2, WNSA-2, WPSA-3, WNSA-3, RPCG, RNCG, RPCS, RNCS) [144], partitioning coefficient (XlogP)

constitutional	Counts of atoms, rings, and bonds (aliphaticAtomCount, aliphaticBondCount, aromaticAtomCount, aromaticBondCount, aromaticRingCount, asymmetricAtomCount, atomCount, bondCount, carboaromaticRingCount, carboRingCount, chainAtomCount, chainBondCount, chiralCenterCount, fusedAliphaticRingCount, fusedAromaticRingCount, heteroaromaticRingCount, heteroRingCount, largestRingSize, ringAtomCount, ringBondCount, ringCount, rotatableBondCount, smallesRingSize), molecular refractivity (refractivity), molecular weight (molWeight), resonant count (resonantCount)	Gravitational indices (GRAV-1, GRAV-2, GRAV-3, GRAV-4, GRAV-5, GRAV-6, GRAVH-1, GRAVH-2, GRAVH-3) [228], Moment of Inertia along the principal axes X, Y, and Z, along with ratios and radius of gyration (MOMIX, MOMIY, MOMIZ, MOMIXY, MOMIXZ, MOMIYZ, MOMIR)
topological	Balaban index (balabanIndex), weighted Burden matrix (BCUTw1l, BCUTw1h, BCUTc1l, BCUTc1h, BCUTp1l, BCUTp1h) [229], Harary index (hararyIndex), hyper Wiener index (hyperWienerIndex), Platt index (plattIndex), Randic index (randicIndex), Szeged index (szegedIndex)	Kier-Hall kappa shape indices (Kier1, Kier2, Kier3), Petitjean number (PetitjeanNumber) and Petitjean indices (topoShape, geomShape), Wiener path number and polarity (WPATH, WPOL), Zagreb index (Zagreb)

Table 14 - Overview of descriptors (n=118) used in this study by origin (ChemAxon: n=58, CDK: n=61) and class.

Software used

ChemAxon Marvin (Marvin 5.0.4, 2008, http://www.chemaxon.com) was used for characterizing chemical structures, substructures, and ChemAxon Calculator Plugins were used for structure property calculation. Additional descriptors were calculated with the open-source cheminformatics package Chemical Development Kit [209] (version 1.0.4, 2008, http://sourceforge.net/projects/cdk). For the calculation of certain descriptors (e.g. the set of charged partial surface area (CPSA) descriptors) 3D structures are required. These structures were generated from SMILES representations of the molecules as given in DrugBank using the Ghemical force field (http://www.uku.fi/~thassine/projects/ghemical/). With this force field, a search for lowest energy conformers was performed using the OpenBabel toolkit (Version 2.2.1, available at http://www.openbabel.org). We used SPSS (version 15.0 for Windows) for DTI and Weka (version 3.4; Waikato Environment for Knowledge Analysis, University of Waikato, Hamilton, NZ,

http://www.cs.waikato.ac.nz/~ml/Weka/) [177] for kNN, ANN, and RF. SVM models were calculated using LIBSVM (version 2.89; http://www.csie.ntu.edu.tw/~cjlin/libsvm). Calculation of chemical diversity and grid screening for SVM meta-parameters was performed with in-house software.

Results and Discussion

Chemical diversity

We calculated the Tanimoto coefficient for the entire data set and arrived at an overall dissimilarity value D(M) of 0.70. This is a high value and confirms the great structural diversity of the compounds studied. Based on this, it is fair to assume a general applicability of our models for the domain of drug-like compounds.

Feature selection

We performed feature reduction for ANN, kNN, and SVM models as described. The 19 descriptors that made the final set were: *asymmetricAtomCount, carboRingCount, dreidingEnergy, fusedAliphaticRingCount, largestRingSize, logD_7, logD_9, rotatableBondCount, BCUTp1h, PNSA1, RPCG, RPCS, THSA, TPSA, Kier3, MOMIZ, MOMIR, WPOL,* and *Zagreb*. This selection seems sensible, as it covers a broad spectrum of chemical features, ranging from simple counts of substructures (e.g. *carboRingCount*) and measures of lipophilicity (e.g. logD descriptors at physiological pH and *TPSA*) to complex topological indices (e.g. *Kier3* and the radius of gyration, *MOMIR*).

Performance for different endpoints

We did not prepare models for inducers for any of the CYP isoforms because of gross underrepresentation of active compounds (Table 13). Instead, we limited ourselves to substrate and inhibitor activities and a combined endpoint (‘global’), for which a compound is labeled as active when it shows any sort of interaction (as substrate, inhibitor, and / or inducer). If no interaction was documented, it was labeled inactive. A summary of CCRs for every endpoint is given in Table 15.

Cytochrome 1A2

Substrates of CYP 1A2 have been noted to be planar, aromatic, and lipophilic compounds with neutral or basic characteristics such as benzopyrene, caffeine, and propanolol [230]. Additionally, Lewis pointed out the relevance of logD partitioning coefficients at physiological pH and hydrogen bond characteristics for substances with ionizable substructures [231]. We see this reflected in the independent variables present in the most effective model (CHAID, CCR: 90.9%): *acceptorSiteCount, asymmetricAtomCount, aromaticRingCount, resonantCount, FNSA3, PetitjeanNumber, dreidingEnergy, fusedRingCount, BCUTp1l, geomShape, rotatableBondCount, logD_5, logD_8, logD_9, logD_13, PSA, aromaticAtomCount, MOMIX, MOMIZ, MOMIXZ,* and *MOMIR.*

Similarly, the global model for any type of CYP 1A2 interaction (CHAID, CCR: 91.5%), emphasizes polar features and geometric properties: *vdwsa, GRAV1, fusedAromaticRingCount, RPCS, FPSA2, logD_9, WNSA3, acceptorCount, carboaromaticRingCount, WPATH, logD_13, WPSA1, logD_7, WPSA2, balabanIndex, DPSA2, logD_0, RHSA, plattIndex, geomShape, DPSA1, BCUTw1h, MOMIZ, molPol,* and *donorCount*. More specifically, five or more hydrogen bond acceptors and at least one hydrogen bond donor appears as a pre-requisite along with the presence of fused aromatic rings.

The best model for inhibition (CHAID, CCR: 81.7%) is markedly less successful. Included were the descriptors *aromaticAtomCount, resonantCount, asymmetricAtomCount, BCUTp1h, logD_74, MOMIYZ, MOMIZ, PetitjeanNumber, topoShape, WNSA1, WPSA2, FPSA1, FPSA3, balabanIndex,* and *WPOL*. The number of active compounds is also lower (half as many as in the substrate class), which suggests that a significant number of structures is classified as false negative.

Cytochrome 2D6

We achieved best results with the DTI algorithms CHAID and CRT and saw similar performance in SVM models. More than in the other endpoints, CYP 2D6 substrate activity seems to rely on ionizability and lipophilicity in the best model (CHAID, CCR: 92.2 %). Descriptors included were *logD_2, logD_3, logD_5, logD_10, logD_14, XLogP, acceptorSiteCount, donorCount, smallestRingSize, heteroRingCount, heteroaromaticRingCount, geomShape, topoShape, PetitjeanNumber, RPCG,* and *FPSA3*. The partitioning coefficients at extreme ends of the pH range (*logD_2, logD_3*, and *logD_14*) are those descriptors which best reflect a compound's acid-base characteristics in the set of features we used. More weight is hence placed on this quality compared to, for example, the CYP 1A2 models presented above. This is in accordance with the common understanding that CYP 2D6 metabolizes its substrates via an ion-pair interaction between basic nitrogen of the substrate and aspartic acid residues at the active site [232].

Performance for the global endpoint is slightly worse, with the CHAID model performing best with a CCR of 89.2%. Relevant descriptors were *acceptorSiteCount, donorSiteCount, aromaticRingCount, asymmetricAtomCount, carboaromaticRingCount, smallestRingSize, logD_1, logD_5, logD_14, PPSA3,* and *PNSA-1*. Generally, basic compounds with aromatic rings interact with CYP2D6 while hydrogen bonding behavior is of lesser importance.

The CRT model for inhibitors of CYP 2D6 had the best performance of any model in this study, relying on the following descriptors: *RHSA, WNSA-1, logD_4, logD_5, logD_11, logD_12, logD_15, FPSA3, dreidingEnergy, donorCount, DPSA1, topoShape, PNSA2, BCUTc1l, and MOMIXY*. Again, acid-base characteristics and lipophilicity appear, but shape constraints are of importance as well.

Cytochrome 3A4

The base for CYP 3A4's broad substrate specificity is thought to lie in its large active site where weak hydrophobic interactions determine binding [230, 233]. Yap and Chen [227] reported on good models using descriptors of shape and connectivity, electronegativity, hydrophobicity, and polarizability. These were not only present in the features selected for SVM and ANN learning but also in the more successful DTI models.

For the global endpoint, we saw best results with the CRT algorithm (CCR: 87.4 %) using *hararyIndex, XlogP, logD_6, logD_11, logD_12, balabanIndex, BCUTp1h, BCUTw1l, RNCG, RPCG, geomShape, WPOL, resonantCount, THSA, aliphaticBondCount,* and *MOMIX*. This was also the case for substrates (CCR: 89.8 %) with *MOMIR, XlogP, logD_0, logD_3, logD_6, logD_74, logD_8, logD_9 GRAV4, MolWeight, WPOL, RPCS, PPSA, refractivity, dreidingEnergy, geomShape, DPSA3,* and *aromaticAtomCount*. Although these CCRs are quite good, they trail behind the values achieved for CYP1A2 and CYP2D6, whose specificity is narrower and hence more easily modeled.

For inhibitors, the CHAID algorithm produced by far the most accurate model (CCR: 87.6 %), using the descriptors: *molPol, GRAV1, PPSA3, WNSA3, GRAVH1, MOMIXZ, WNSA1, logD_0, logD_1, logD_12, logD_13, plattIndex, PSA, acceptorSiteCount, BCUTw1l, PNSA2, resonantCount, carboRingCount, PPSA1,* and *rotatableBondCount*.

Comparison of ML methods

In all endpoints, we achieved the best results with DTI. At first glance, this seems to contradict studies such as Vasanthanathan et al. [226], whose modeling of CYP 1A2 inhibitor activity for approximately 7400 substances using a similar portfolio of methods showed highest accuracy with SVM. DTI, however, often outperforms numerical methods (e.g. SVM or ANN) for smaller data sets, even with prior feature selection [29]. Other groups have reported accuracies of well over 90 % correctly predicted instances for SVM models [226, 234], but only when applied to training sets. The same models achieve significantly lower predictive power in 5-fold CV (around 75 to 80 %).

Conclusions

We present a dataset of 353 compounds and their interaction status with CYP isoforms 1A2, 2D6, and 3A4. Additionally, we give cross-validated QSAR models for these activities using a set of common ML methods. Of these methods, DTI consistently outperforms its competitors. This is most probably due to the usefulness of DTI algorithms for smaller datasets.

In our survey of over 1'400 FDA approved compounds in DrugBank, the most frequent interactions were with CYP 3A4. In literature, this enzyme is implicated in the metabolism of more than 50% of drug compounds [235], so the high proportion of CYP 3A4 active compounds (n=264, 74.8%) can be expected.

However, these account for only about 18% of the compounds in DrugBank. We therefore assume that additional screening of these could uncover many other drugs with CYP 3A4 interaction potential. Arguably, these interactions could be of lesser clinical importance, either because the offending drug is not as widely used or the interaction potential is less poignant.

Furthermore, comparing our combined activity endpoints (any sort of interaction as substrate, inhibitor, and / or inducer) with the isolated activities substrate and inhibitor for each isoform, we would have expected to see significantly better CCRs in the latter, especially when the number of active compounds approaches those of the combined endpoint. This was seldom the case but is easily explained by the parallel membership of many compounds in different classes. For example, the anti-hypertensive agent metoprolol and the first generation histamine receptor antagonist promethazine both act as CYP 2D6 substrates and inhibitors, while the calcium channel blockers verapamil and amlodipine have both activities in CYP 3A4. There is significant overlap in substances and therefore in characteristics relevant for activity. In the case of CYP 1A2, models for inhibitory compounds performed markedly worse than for the combined endpoint, although the decrease in active substances is the same range as for the other isoforms. Most probably this is due to a large number of false negative instances in the dataset.

With CCRs of 81.7 to 92.9 %, our models are well suited to guide compound selection in drug discovery and also to differentiate between interactions between the different isoforms studied rather than CYP interactions in general. This is echoed in the differences between descriptors selected by the DTI algorithms for the respective endpoints. Models do share certain general features of drug likeness (such as lipophilicity and ionizing behavior) but rarely make use of very general descriptors such as molecular weight or measures of connectivity (e.g. the Kier and Hall kappa shape indices). The latter is often the case in QSAR studies of mass screenings of large chemical libraries, where algorithms must establish drug likeness in addition to the actual pharmacological or toxicological endpoint itself.

Our analysis focuses on easily calculatable descriptors and freely available modeling tools. While predictive accuracies are very high, further improvements could be expected from more sophisticated methods. Highly resolved crystal structures have been published for many CYP P_{450} isoforms which have special relevance to drug metabolism [236], allowing the use of target-based approaches. Other groups have used methods similar to ours. In particular, Yap et al. have presented work on CYP P_{450} interactions with statistical learning methods [227, 237] as have Leong et al. [238, 239]. Their studies show a slightly higher predictive accuracy for SVM approaches, esp. when augmented with pharmacophore information. DTI models of CYP 3A4 inhibitors [237] were of lesser accuracy than the ones we present. However, most studies make use of the C4.5 algorithm [192] which is often outperformed by algorithms like the ones we employed. [240] This illustrates two important facets of QSAR modeling. Firstly, no single method gives the best results for every dataset. Secondly, the success of a given method is dependent on subtleties of application (setting of learning parameters, preferences of algorithms employed, etc.).

	CYP 1A2			CYP 2D6			CYP 3A4		
	global	substrates	inhibitors	global	substrates	inhibitors	global	substrates	inhibitors
RF	66.7	57.3	61.9	78.1	76.3	72.3	67.5	73.0	65.8
kNN	69.7	56.6	64.3	79.0	77.2	76.1	62.3	72.4	64.2
ANN	67.4	63.2	57.3	79.6	70.9	75.4	61.9	67.4	62.8
CHAID	**91.5**	**90.9**	**81.7**	**89.2**	**92.2**	91.6	81.4	88.6	**87.6**
CRT	78.2	78.6	70.9	87.0	89.4	**92.9**	**87.4**	**89.8**	84.7
SVM RBF	71.2	66.0	66.6	83.1	76.5	77.4	67.2	66.0	66.4
SVM polynomial	68.8	63.3	63.0	82.6	75.8	74.4	66.0	65.1	62.9

Table 15 – Corrected classification rates of 10-fold crossvalidated QSAR models of cytochrome 1A2, 2D6, and 3A4 interaction. RF: Random Forest, kNN: k-nearest neighbors, ANN: artificial neural networks, CHAID: Chi-squared interaction detector, CRT: classification and regression trees, SVM RBF: support vector machines using the radial basis function kernel, SVM poly: support vector machine using the homogeneous polynomial kernel. Best values are in bold-face.

5.1.3. Prediction of Adverse Drug Reactions using Decision Tree Induction

Felix Hammann [1], Heike Gutmann [1§], Nadine Vogt [1], Andreas Maunz [2], Christoph Helma [2], Juergen Drewe [1*].

[1]: Department of Gastroenterology & Hepatology, University Hospital Basel, University of Basel, Switzerland

[2]: Freiburg Center for Data Analysis and Modelling, Albert-Ludwigs-University Freiburg, Germany

[§]: present address: Novartis Institutes for Biomedical Research, Drug Metabolism and Pharmacokinetics, Basel, Switzerland

.

Corresponding Author (*):

Prof. Dr. Juergen Drewe

Department of Gastroenterology & Hepatology

University Hospital of Basel

Petersgraben 4

CH-4031 Basel Switzerland

Email: juergen.drewe@unibas.ch

Phone: +41-61-265 3848

Fax: +41-61-265 8581

Clinical Pharmacology and Therapeutics (accepted)

Abstract

Drug safety is of great importance to public health. Detrimental effects of drugs not only limit their application, but also cause suffering in individual patients and distrust of pharmacotherapy. To identify suspicious drugs, we present a structure-activity relationship analysis of adverse drug reactions (ADRs) in the central nervous system (CNS), liver, kidney, and for allergic reactions for a broad variety of drugs (n=507) from the Swiss drug registry. Using decision tree induction, a machine learning method, we determined predisposing chemical, physical, and structural requirements. The models had high predictive accuracies of 78.9% to 90.2% for allergic, renal, CNS, and hepatic ADRs.

We show the feasibility of predicting complex end-organ effects using simple and computationally inexpensive models, which can be used 1. in compound selection in drug discovery, 2. to understand how drugs interact with the target organ systems, and 3. for the generation of alerts in post-marketing drug surveillance and pharmacovigilance.

Keywords: QSAR, drug safety, toxicology, central nervous system, kidney, liver, allergies, adverse drug reactions, ADR

Introduction

No drug has only beneficial effects for the patient. Adverse drug reactions (ADRs) are undesirable effects that occur when a drug is administered at the proper dose in the correct manner for an appropriate indication, i.e. the drug was used correctly and in good faith. ADRs are part of the larger group of adverse drug events (ADEs) which encompass any noxious effects related to drug therapy. ADEs therefore also include drug overdose, drug-drug-interactions, and other medication and prescription errors. Strategies to reduce ADEs due to (human) error are more easily devised than for ADRs. Here, the burden lies mostly with the pharmaceutical industry to develop safe compounds and test them thoroughly. Pharmacovigilance increases the knowledge of ADRs in daily practice, especially for less frequent adverse events. This requires a larger population to be medicated before effects can be observed. While ADRs may not be preventable they may be anticipated.

The avoidance and proper management of ADRs are of great importance to public health, and the public at large shows high interest in drug safety. Unrecognized or underreported ADRs not only cause preventable human suffering and costs to the health care system, but can also unnecessarily undermine the public's faith in (and compliance with) drug therapy. This issue was addressed by the 110th United States Congress which passed an act in 2007 to increase post-marketing surveillance of regulated drugs. The FDA has since started the implementation of the so-called Sentinel Initiative, integrating data from health-care providers, post-market surveillance agencies, and other sources.

The technical pre-requisites of monitoring ADRs and timely generation of alerts are essentially those of data mining (finding and extracting patterns from potentially huge sources of data using mathematical and statistical methods). It is important to have analytical tools that are simple and robust, yet specific. For drug safety assessments, an optimal system would receive structural information of the drug and produce notifications that either rule out any noxious potential or point specifically to ADRs that could be expected to arise. Ideally, such a system would arrive at its notifications in a straight-forward way, rather than breaking the problem down into different smaller tests (e.g. detailed physiological modeling of all processes from drug ingestion to the target structures).

Data sources for the prediction of clinical effects

Many *in vitro* and animal models exist to screen for toxicological effects, but these often do not translate well into clinical practice. Although *in vitro* assays allow activity and ligand optimization, the environments within which they are performed rarely conform to conditions in human organisms. Also, solubility and permeability can differ greatly between animals and humans because of different acidity and transit time in animal gastrointestinal tracts [160]. Lastly, ADRs which develop only under chronic use or which impair specific human functions (such as higher cognitive functions) are not readily evaluated by these models. For these reasons, human *in vivo* data are preferable.

Such extensive safety data are not easily obtained. A great and largely untapped source, however, lies in the information collected by regulatory offices and pharmacovigilance sites in the form of ADR reports. For this study, we aim to create a comprehensive survey of ADR reports for a broad variety of drugs in clinical use and develop computational models for understanding and predicting such reactions. Basis for predictions are numerical features easily calculated from nothing but the chemical structure.

Classification of ADRs

A drug must first reach an organ to exert an effect, be it beneficial or not. For central nervous system (CNS) effects, a common paradigm therefore holds that drugs must be able to cross the blood-brain barrier (BBB). This is certainly true for substances that act directly on target structures in the CNS, but ADRs such as encephalopathy or delirious states can just as well be caused indirectly, for example, by hepatotoxic or nephrotoxic compounds which induce liver or kidney failure. Most neurological ADRs can be strictly assigned to the CNS. With psychiatric reactions, this is not as easy. Dependency and addiction, for example, can be caused by modulation of central receptors (in the case of opioid addiction). However, they may also be of psychological nature or due to peripheral effects (e.g. dependency on vasoconstrictor nasal sprays). For this study, we have therefore chosen to label an ADR as CNS-related only if its origin is at least predominantly within the CNS.

Drug-induced liver injury is another major complicating factor in drug therapy and an important reason for withdrawals from the market, especially in the case of idiosyncratic reactions [241]. Risk factors for hepatotoxic ADRs (diabetes, obesity, old age) [242] are common and will gain in importance as the

average lifespan continues to increase. Compromised liver function results in decreased clearance of hepatically metabolized substances and toxic compounds which eventually accumulate. Ultimately, hepatic encephalopathy may develop, owing to the accumulation of ammonia and mercaptans. In this way, hepatotoxic ADRs can have CNS effects by proxy and should be taken into consideration in the evaluation of a drug's CNS safety.

The same holds true for renal ADRs. The kidney is central to the elimination of water-soluble metabolites and xenobiotics. Renal failure can lead to accumulation of primarily nitrogenous compounds (uremia) with devastating effects on the CNS. Nephrotoxic ADRs are seen in cardiovascular drugs (esp. ACE inhibitors), aminoglycosides and NSAIDs (direct tubular damage and interstitial nephritis), and a broad range of other drugs.

Lastly, we consider allergic ADRs. These form a group of commonly reported problems, ranging from laboratory finding (e.g. eosinophilia) to local irritations or skin manifestations over to very serious complications such as first manifestations or exacerbation of allergic asthma. Allergic ADRs are therefore also dependent on a patient's medical history and individual susceptibility. Events attributable to medication errors (esp. those which are document only for overdose) were not included.

Quantitative Structure-Activity Relationships (QSAR)

See Section 3.1.2.

For this study, we restricted ourselves to decision tree induction analysis. Such models not only perform as prediction tools but also lend themselves to visual interpretation and provide insight into the mechanisms behind the endpoints.

Materials and Methods

Data acquisition

We based our analysis on a list of FDA approved small molecule drugs (n = 1436, date of access: January 1st, 2008) from the University of Alberta's DrugBank database [225]. Currently, DrugBank contains over 1'300 FDA approved small molecule drugs along with chemical, pharmaceutical and biochemical information. Using the official Swiss drug registry, we retrieved a list of every ADR (n = 776) known for the subset of 507 compounds in DrugBank which are also marketed in Switzerland. Every individual ADR was labeled with one or more of the following tags: CNS related, hepatotoxicity related, nephrotoxicity related, allergy related, and miscellaneous. ADRs which could only be classified as 'miscellaneous' were not considered.

Unspecific ADRs such as fatigue, nausea, headache, and feelings of dizziness were removed from the final dataset because these reactions are almost regularly seen in clinical trials as a stress reaction to the

unfamiliar setting and rarely verified by an external examiner. We also removed dependency, anorexia (as a consequence of nausea), and very rare or insufficiently characterized ADRs such as fever (unless explicitly stated as drug fever; n = 35), symptoms of abstinence (n = 26), and death (n = 9). Based on the class labels, we computed the sums of ADRs for every class and every compound.

For every ADR class, compounds at a low and a high risk were defined to have ADR counts equal to or lower than a preset lower threshold or equal to or higher than a preset upper threshold, respectively. Compounds with ADR counts between the two thresholds were excluded from the analysis for that class. The concept of using these experimentally determined thresholds allowed to obtain sharper distinctions between low-risk and high-risk groups of compounds. A summary of thresholds used is given in Table 16. A table of structures and activities in each class can be found as supplemental material.

class	lower threshold	upper threshold	active compounds	inactive compounds	compounds of undetermined activity
CNS	1	20	97	67	343
hepatotoxicity	0	4	109	177	221
nephrotoxicity	0	3	126	208	173
allergies	0	2	99	239	169

Table 16 – Description of the dataset used (n = 507). Adverse drug reactions (ADRs) are grouped into one or more of the given classes. Compounds with an ADR count ≤ lower threshold were classified as inactive in the respective class, those with an ADR count ≥ upper threshold were classified as active. All others were classified as undetermined.

Assessment of chemical diversity

See Section 4.1.6.2.

Decision Tree Induction (DTI)

See Section 4.2.4.

Corrected Classification Rate (CCR)

See Section 4.2.2.1.

Model Validation

See Section 4.2.2.5.

Descriptors

See Section 4.1.7.

Software used

ChemAxon Marvin (Marvin 5.0.4, 2008, http://www.chemaxon.com) was used for characterizing chemical structures, substructures, and ChemAxon Calculator Plugins were used for structure property calculation. Additional descriptors were calculated with the open-source cheminformatics package Chemical Development Kit [209] (version 1.0.4, 2008, http://sourceforge.net/projects/cdk). Decision trees grown with the CHAID and CART algorithms were done in SPSS 15.0 for Microsoft Windows.

Results and Discussion

We calculated the dissimilarity for all compounds in the data set with the Tanimoto coefficient and arrived at a value of 0.85. This rather high value is evidence of our data's great structural diversity.

Overall, we found very few substances without any CNS ADRs. However, when ADRs have a psychiatric or psychosomatic character, they become both very subjective and harder to evaluate for non-specialists. Furthermore, a great deal of ADR data is acquired in the form of adverse events during phase I – III clinical trials. Under these artificial conditions, volunteers and patients are confined to trial units and/or constantly monitored. This induces significant stress and oftentimes finds expression in 'soft' ADRs. Although these may very well be attributable to the study drug, the relationship is seldom clear.

The safest substances in our set judging by the total number of ADRs were the vasodilator and anti-alopecia agent minoxidil (cardiovascular and pulmonary ADRs are known) and lactulose (gastro-intestinal but not hepatotoxic ADRs). The highest ADR potential was achieved by the HIV-protease inhibitor ritonavir, the diabenzazepine carbamazepine, and the calcineurin inhibitor tacrolimus.

Performance for different endpoints

A summary of CCRs for every endpoint and DTI algorithm is given in Table 17.

	CNS	Liver	Kidney	Allergies
CHAID	**89.74 %**	87.15 %	84.71 %	78.46 %
CRT	88.01 %	**90.22 %**	**88.69 %**	**78.94 %**

Table 17 – Corrected classification rates for the different classes of adverse drug reactions by decision tree induction algorithm used. Best values are in boldface.

CNS class ADRs

A decision tree generated with the CHAID algorithm performed very well at 89.7 % accuracy (Figure 50). Specifically, compounds with less than 10 chain atoms, no carboxyl groups, one or no amine groups, absence of azoles, a partial polar surface area (PPSA-1) less than 293.7 $Å^2$, and logP below 3.1 show the least CNS ADRs. Except for the low lipophilicity requirement, these are all characteristics of drugs with good BBB permeability. This implies that these compounds are not safe merely because they do not primarily cross the BBB. A possible explanation is that compounds are extruded by efflux mechanisms. Although activation of efflux pumps can give rise to drug-drug interactions (and, by consequence, be the source of medication errors and ADEs), drug efflux effectively prevents brain penetration and potentially toxic concentrations. One of the major efflux pumps of the BBB, P-glycoprotein (P-gp, MDR1, ABCB1), was the subject of a recent study conducted in our laboratory [243]. We therefore cross-checked the compounds in both databases and found 49 matches. Only 12 of 25 compounds with CNS ADRs (48 %) and 4 of 8 compounds without these ADRs (50 %) are reported to be P-gp substrates, indicating that clearance from the CNS is no sufficient explanation for safety. Rather, it seems that the model reflects the important goals of good permeability and low number of ADRs in drug discovery. This view is further supported by the requirement for moderate to low lipophilicity, which places a limit on accumulation in the fatty matrix and membranes of the CNS. The CNS concentrations of such compounds are easier to manage and predict, especially in chronic use. Minoxidil was one of the few compounds without any CNS ADRs in our dataset and is a good example in this regard.

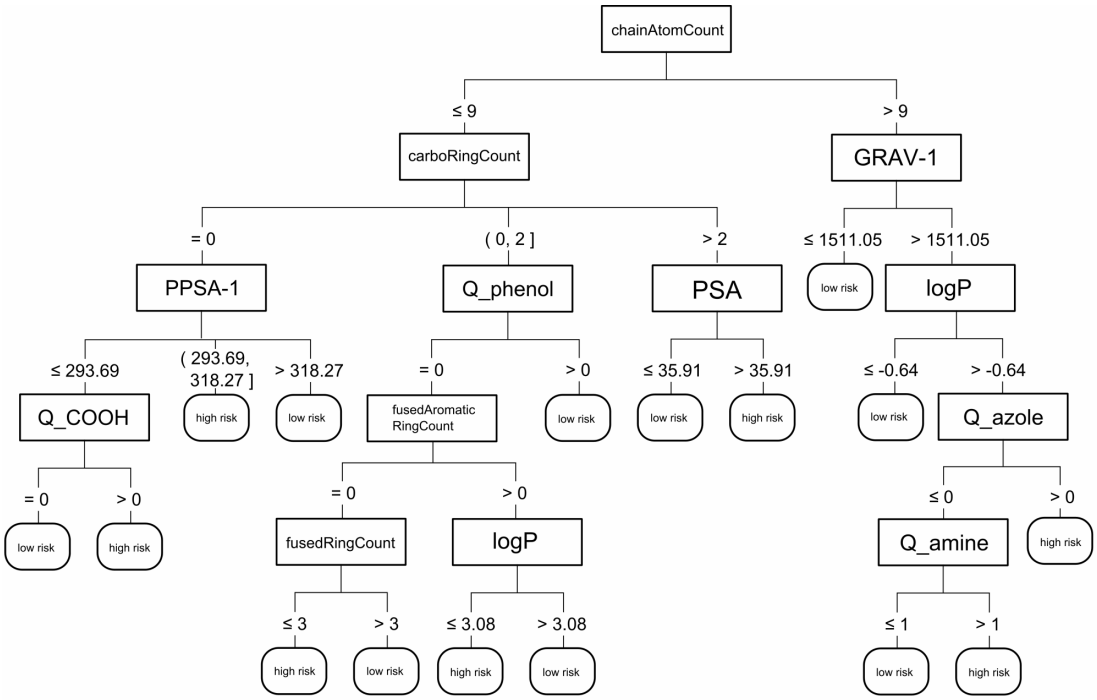


Figure 50 - CHAID decision tree model for central nervous system class adverse drug reactions for 97 active and 67 inactive compounds (n = 164) with a corrected classification rate of 89.74 %.

Hepatotoxic class ADRs

Best performance was seen with a CART decision tree model (CCR: 90.2 %, Figure 51). The model is quite complex, spanning 39 nodes with a total depth of 5. Most prominently featured are descriptors of charge and lipophilicity. Important characteristics of safe compounds were a small partial negative surface area (PNSA-1 < 167.72 $Å^2$), small atom-weighted partial positive surface area (PPSA-3 < 35.16), small polar surface area (PSA < 94.88 $Å^2$), low nitrogen count (less than 3). Interestingly, the counts of important substructures such as azoles or nitro groups were not deemed relevant in the model. The CHAID decision tree model (CCR: 87.15 %) showed similar descriptors, but in addition found low sulfur atom and carboxyl and amine group counts, as well as a hydrogen bond donor site count of less than four, and places importance on low lipophilic partitioning. Small molecules such as pyridoxine, a form of vitamin B6 given to lessen the toxic effects of the isoniazide, and the histamine H_1 receptor antagonist meclizine used in treatment of motion sickness, vertigo, and nausea during pregnancy, are examples of substances with small hepatic ADR potential.

This is in line with the common understanding that hepatotoxicity is achieved mostly by lipophilic compounds, which not only have higher permeability but also accumulate more easily. The substructure motifs identified by the CHAID algorithm are the basis of toxic metabolites. It seems wise to avoid compounds which can be subject to metabolic activation and thus be the source of noxious electrophilic species [244]. Ritonavir, for example, is rich in nitrogen, has a large PSA and PNSA-1, as is atazanavir. Both are lipophilic and noted for their detrimental effects on the liver.

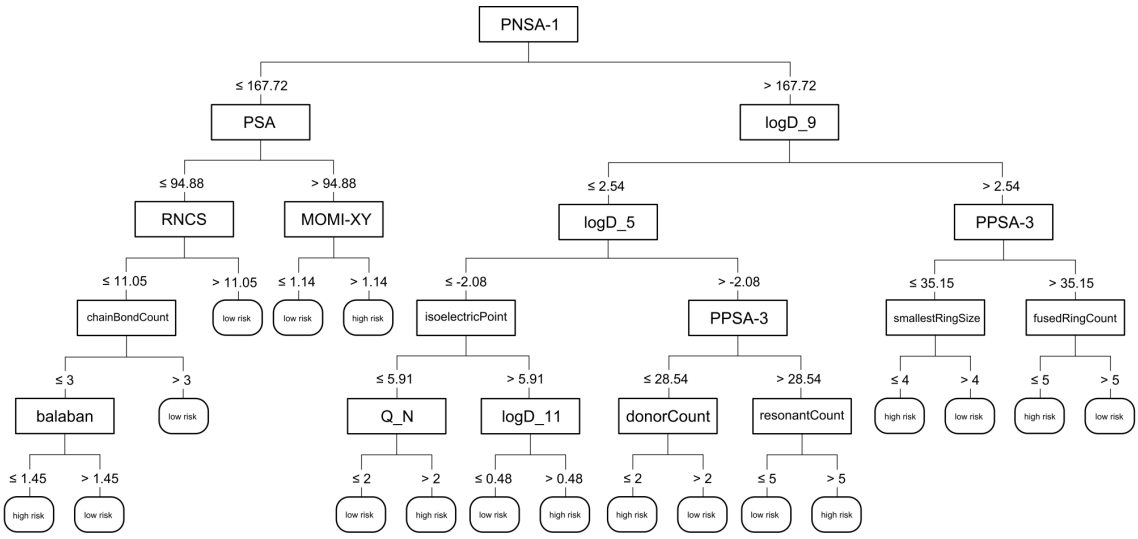


Figure 51 - CART decision tree model for liver class adverse drug reactions for 109 active and 177 inactive compounds (n = 286) with a corrected classification rate of 90.22 %.

Nephrotoxic class ADRs

We saw very good performance with the CHAID and CART (Figure 52) decision tree models (CCRs: 84.7 and 88.6%). Again, low PSA (< 93.78 $Å^2$) is an important factor for safety. Few aromatic atoms (less than 19), a basic pKa under 10.71, van der Waals surface area (vdwsa) under 1014.5 $Å^2$, and logP values over 2.43 were seen in the CHAID model for safer substances. The CART model additionally points out the detrimental influence of amine functions, sulfur, and carboaromatic ring structures.

Despite the comparatively high threshold for van der Waals surface area (vdwsa) the models associate lipophilic and non-polar compounds with good renal safety. Such compounds are more likely to undergo hepatic rather than renal elimination, and, by consequence, will not have great exposure to renal parenchyma. Tacrolimus appeared as one of the least safe compounds with regard to kidney and liver related ADRs. This is well known [245] and it has many structural features that are looked for in our models (high van der Waals surface area, nitrogen count, and logP).

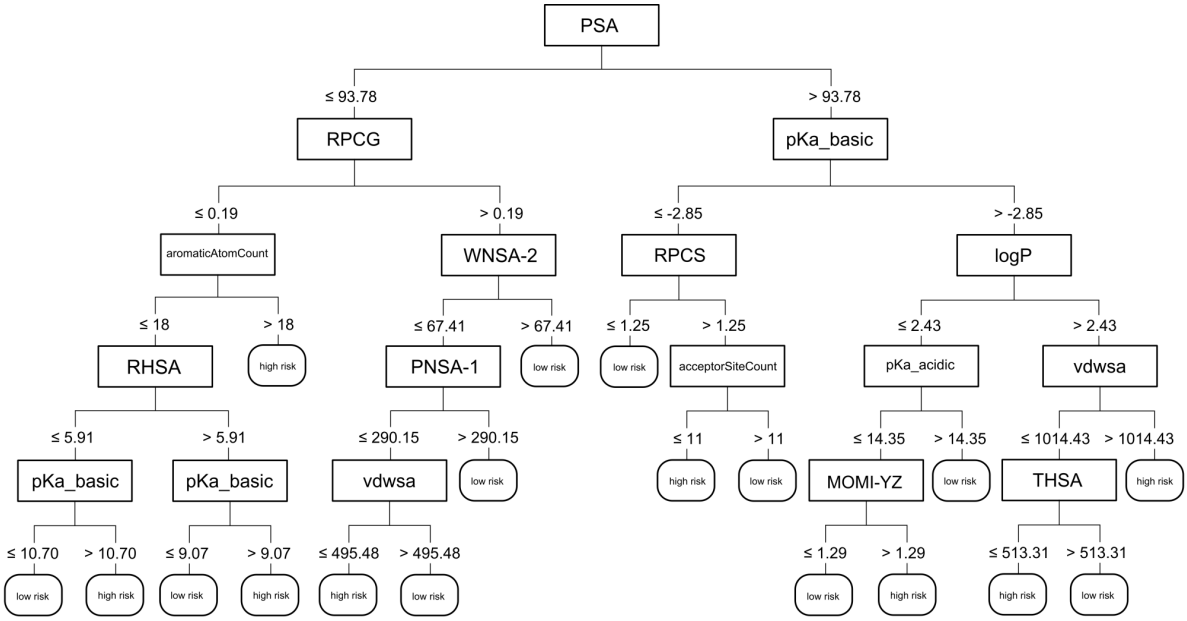


Figure 52 - CART decision tree model for kidney class adverse drug reactions for 126 active and 208 inactive compounds (n = 334) with a corrected classification rate of 88.69 %.

Allergy class ADRs

Compared to the other endpoints, CCRs for this class were lower. Satisfactory models at almost equal CCRs were provided by the decision tree algorithms (CHAID: 78.4 %, CART: 78.9 %). A diagram of the CART model is given in Figure 53. Both also place emphasis on topological complexity over the mostly permeability and substructure oriented models above. Of note are descriptors such as the third Kier kappa shape index (which relates structural complexity to extreme shapes of chemical graph theory), the Wiener

index (a measure of the degree of branching), and relative moments of inertia, i.e. MOMI-XY and MOMI-XZ, reflecting the degree of symmetry.

Still, the predictive accuracy is somewhat impressive, given the great structural diversity of compounds with antigenic potential. It is likely precisely this circumstance to which these models owe their performance. Instead of basing their predictions on known antigenic motifs (which were largely not determined by the descriptor set used), both models rely on abstract measures of complexity and state that simple, less branched compounds are safer. This is in line with common immunological understanding, where complex proteins are far more immunogenic than polysaccharides. Also, smaller and less complex molecules do not allow for the repetitiveness or structural heterogeneity that triggers an immunological response. The models therefore support the observation that very sophisticated drug molecules such as monoclonal antibodies are potential sources of allergic ADRs.

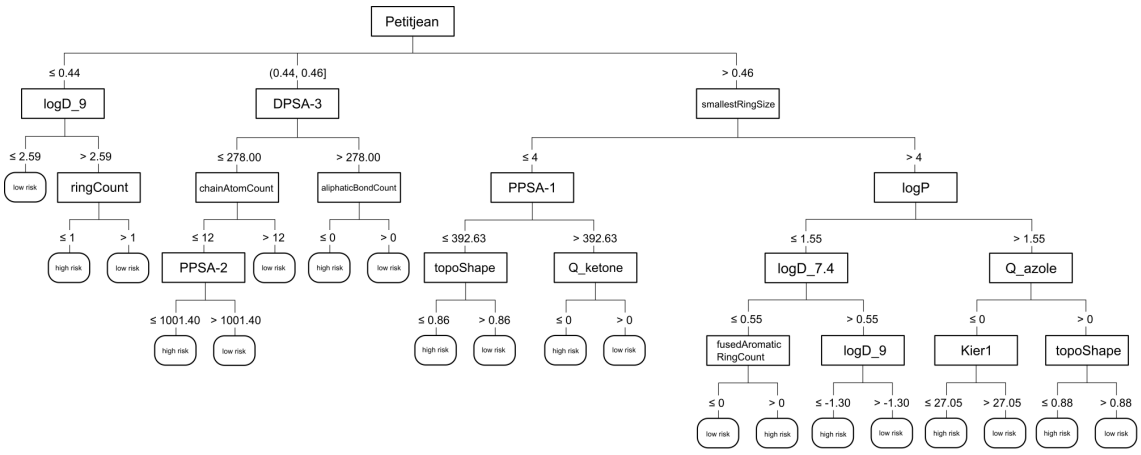


Figure 53 - CART decision tree model for allergy class adverse drug reactions for 99 active and 239 inactive compounds (n = 338) with a corrected classification rate of 78.94 %.

Conclusions

In this study, we performed QSAR analyses using DTI for compounds causing several classes of ADRs. ADRs may also be considered as a surrogate marker for a compound's ability to interact with the target organ *in vivo* and in humans (data which is notoriously hard to acquire), and all models use very different features to arrive at their decisions. This suggests that they do not simply model general characteristics of drug-likeness but specifically their endpoints. Our models can therefore be directly used as screening tools in late drug discovery as structural guidelines, in rational drug design, and pharmacovigilance evaluations. However, it should be noted that these models are only valid for drug-like molecules and are not necessarily suitable for other areas such as environmental toxicology.

CNS ADRs were reported for almost every drug in our list. This is not surprising, as the assessment of CNS symptoms is often subjective to a patient's self-perception and a clinician's evaluation of it. They are

also much more influenced by external factors than are the other ADR classes examined here. Despite this, we were able to provide a thorough estimation tool for this major complicating factor in drug therapy. Models for hepatic ADRs rely on lipophilicity and count substructures. Considering the metabolic role of the liver in bioactivation by unmasking, modifying, and conjugating certain substructures, this is not surprising. Given the complexities of hepatic metabolism, it may be hard to significantly improve on the performance of general hepatic ADR models such as the ones we present. The models for renal ADRs suggest that safer compounds are less permeable and might not reach sufficiently high concentration in the renal parenchyma. Of course, many of the compounds are at least partially eliminated by the kidney, but often substances need to undergo prior metabolic alteration. Better accuracy could therefore be achieved if this is taken into account. However, ADR data as used in this study does not permit such an analysis. Finally, we presented very consistent (although only satisfactorily performing) models for allergic ADRs, relying on abstract measures of compound complexity. The results are in line with current understanding of antigenicity.

We would like to point out that, while our findings are practically applicable, the requirements presented are more of a disjoint set of rules than a final theory of ADRs. Specific features discussed are used within the models to arrive at a decision and neither necessarily need to be all satisfied for a given compound nor does one necessarily suffice. Furthermore, we have presented rather broad classes of ADRs rather than models for more specific effects. While this pooling was necessary to obtain robust and expressive models based on a sufficient number of compounds, it also reflects how ADRs appear in clinical practice. Rather than causing a single detrimental effect in one organ system alone, multiple related effects are the rule rather than the exception.

The different endpoints presented here are traditionally seen as the sum of different transport, metabolism, and docking events. For example, for direct CNS toxicity to occur in a peripherally administered drug, a compound must appear in the peripheral circulation, pass through any of the blood-CNS-barriers (most notably, the BBB), face potential active efflux, and be able to interact with a central binding site. While it is mechanistically appealing to trace this path by a series of physiological models for these events, it must be noted that each prediction is made with a varying degree of uncertainty (which cumulates with an increasing number of models put in sequence). The analyses in this paper, however, prove that it is in fact possible to model such very complex phenomena with simplistic means.

Additionally, our models require very little computational power and could easily be applied in high-throughput virtual screenings such as required in post-market surveillance. Uncertainty is always associated with the use of predictive models such as the ones presented here. Although our models are built on data of drug effects in humans (rather than *in vitro* assays of very specific toxicological mechanisms), they are applicable only to small molecule drugs like those encountered in clinical practice. Our models are meant to guide the decision-making process of professionals in the pharmaceutical industry, clinicians involved in pharmacological therapy, and other experts.

Instead of developing models for individual ADRs, we chose to combine related ADRs to form the larger classes described above. It would certainly be desirable to have predictive tools for individual ADRs, esp. when it comes to serious or potentially fatal entities. However, ADRs (n = 789) outnumber available compounds (n = 507) considerably and modeling of every entity is therefore impractical. Also, specific toxicological effects may be better modeled by analysis of culprit targets (e.g. interactions with the human ether-à-go-go related channel (hERG) to screen for QT prolongation).

Different analytical tools might also be applied to our dataset. Linear models (e.g. multiple linear regression), artificial neural networks, support vector machines, or any other of the currently favored tools may yield additional information or improved models. Our desire, however, was to give easily interpretable and practically applicable models (a strong point of DTI) along with a principal proof of feasibility.

5.1.4. Determination of the Single Nucleotide Polymorphisms C3435T and G2677T in MDR1 and C421A in BCRP in blood samples of patients with Inflammatory Bowel Disease and healthy controls in the Swiss population

Felix Hammann [1], Heike Gutmann [1§], Petr Hruz [1], Jyrki Eloranta [2], Stephan Vavricka [2], Gerd Kullak-Ublick [2], Jürgen Drewe [1]

[1]: Department of Gastroenterology & Hepatology, University Hospital Basel, University of Basel, Switzerland

[2]: Department of Clinical Pharmacology & Toxicology, University Hospital Zurich, University of Zurich, Switzerland

[§]: present address: Novartis Institutes for Biomedical Research, Drug Metabolism and Pharmacokinetics, Basel, Switzerland

Corresponding author:

Prof. Juergen Drewe
Department of Clinical Pharmacology and Toxicology
University Hospital of Basel
Petersgraben 4
CH-4031 Basel Switzerland
Email: juergen.drewe@unibas.ch
Phone: +41-61-265 3848
Fax: +41-61-265 8581

Abstract

Aims: The efflux pumps P-glycoprotein (P-gp, ABCB1, MDR1) and Breast Cancer Resistance Protein (BCRP, ABCG2) protect the luminal cells of the gastro-intestinal tract from potentially toxic substances. Genetic polymorphisms of these proteins have previously been associated with disease susceptibility, disease severity, and treatment prognosis of inflammatory bowel diseases such as Crohn's disease (CD) and ulcerative colitis (UC). In this study, we investigated the prevalence in the Swiss population of frequent single nucleotide polymorphisms of P-gp and BCRP in healthy volunteers (n = 17) and patients newly diagnosed with CD (n = 34) or UC (n = 38). **Methods:** We isolated DNA from peripheral blood cells and assessed the genotype and allele frequencies of MDR1 C3435T, MDR1 G2677T, and BCRP C421A using allelic discrimination assays (SNP genotyping; TaqMan). **Results:** We saw weak associations for BCRP C421A ($p < 0.18$) and MDR1 G2677T ($p < 0.27$) in patients with UC compared with healthy controls and a trend towards the wild type allele for MDR1 C3435T ($p < 0.46$) in UC. In haplotype analysis, MDR1 3435CC / BCRP 421CC (X^2: 1.0142, $p < 0.30$) in UC and MDR1 2677G / BCRP 421A (X^2: 1.5615, $p < 0.22$), also in UC, showed the strongest correlations. Results for BCRP C421A in particular justify further study.

Introduction

Inflammatory Bowel Diseases (IBDs) are a group of high-incidence inflammatory illnesses of the intestine, the two most prominent of which are ulcerative colitis (UC) and Crohn's disease (CD) [246]. Both share key clinical features (e.g., nausea, maldigestion and malnutrition, and associated extra-intestinal manifestations [247]) but also differ in their location and the type of inflammation (UC is restricted to the mucosa of the colon whereas CD affects the whole length of the gastrointestinal system and all three layers of the epithelium).

There is increasing evidence that the common cause of IBDs is a malfunction of the intestinal immune system and screening efforts support the existence of genetical susceptibility loci [248] which also play a role in treatment options and prognosis [249]

P-glycoprotein

See Section 3.3.1.3.

Breast Cancer Resistance Protein (BCRP)

See Section 3.3.1.4.

Synergisms of P-gp and BCRP

ABC transporters, and P-gp and BCRP in particular, play an important role in tissue defense. Both are highly expressed in so-called side population (SP) cells, a line of primitive cells derived from bone marrow stem cells. SP cells have been detected in non-hematopoietic tissue and the two efflux pumps are thought to contribute strongly to tissue defense and regeneration in many organs [250]. A further synergism between P-gp and BCRP lies in their shared substrate affinity, for example the anti-cancer agent doxorubicin [251], the α_1-receptor antagonist alfuzosin, and the histamine H2-receptor antagonist cimetidine [252]. Also, recent work by Tai et al. [253] indicates that both are involved in the clearance of the neurotoxic amyloid beta from the brain, thereby protecting individuals from Alzheimer's disease. It therefore seems wise to consider these two efflux pumps together in studies of pharmacokinetics or pathogenesis.

Haplotype studies

Haplotypes of MDR1 and BCRP have been reported to be associated with different IBDs. Urcelay et al., for example, have identified a susceptibility haplotype (2677T / G3435) for CD [254]. Fiedler et al. [255] found a similar association, most notably for the MDR1 2677GG / 3435TT haplotype with UC, but not for CD. The haplotypes investigated by Ho et al. [256] showed positive (MDR1 G2677 / 3435T) and negative (MDR1 2677T / C3435) correlations with UC, and confirmed the findings of Fiedler et al. Haplotypes including BCRP C421A have been less frequently assessed and no associations with IBD are known of to date [257].

Methods and Materials

Subjects

We recruited a total of 89 unrelated volunteers (54 female, 35 male; age: 49.9 years ± 0.2 SEM; weight: 72.3 kg ± 0.2 SEM; height: 1.68 m ± 0.01 SEM; BMI: 25.5 kg/m^2 ± 0.1 SEM) in a series of biopsies and blood samples taken between 2001 and 2007. Diagnosis of UC and CD were according to current clinical knowledge and based on radiological, endoscopic, and histopathological assessment [258]. The study was approved by the local ethics committee (Ethik-Kommission Beider Basel, EKBB) and informed consent was obtained by all volunteers.

Genotyping

Genomic DNA was isolated from peripheral EDTA-blood using the QIAamp DNA blood Kit (Quiagen, Hilden, Germany). We performed TaqMan analysis on a 7900HT Sequence Detection System (Applied Biosystems, Rotkreuz, Switzerland). 1µl of 10 ng/ µl of genomic DNA was added in a well for a multiplex allelic discrimination assay along with 9 µl solution consisting of Applied Biosystems TaqMan MasterMix, forward primer, reverse primer, probes specific for the SNP being examined (see below), and RNase-free

water. Samples were pipette on a 384-well PCR plate (Treff Lab®). The probe stock solutions (=100 pmol/µl) were diluted to a concentration of 2 pmol/µl for allelic discrimination analysis. During a run, samples were heated to 50°C for 2 minutes, then subjected to 95° C for initial denaturation. Then, 40 cycles of a two-step PCR were performed at 95° C for 15 s and 60° C for 60 s.

Primers and probes

The primer set for BCRP C421A was designed with Primer Express software (Version 2.0, Applied Biosystems) and ordered at Invitrogen (Carlsbad, CA, USA). The sequences for the probes were taken from work by Korenaga et al. [259] (Table 18). For MDR1 G2677T, we used a custom assay ordered from Applied Biosystems. We assessed the MDR1 C3435T SNP with primers and probes according to work by Eap et al. [260].

	sequence
probe 1	5´-FAM-CTGCTGAGAACTGTAAGT-MGB-3´
probe 2	5´-VIC`-CTGCTGAGAACTTTAAGT-MGB-3´
forward primer	5'-TGTTGTGATGGGCACTCTGAC-3'
reverse primer	5`-TCATAGTTGTTGCAAGCCGAA-3`
Artificial templates	
positive control (C allele)	5'-TCATAGTTGTTGCAAGCCGAACTGCT GAGAACT**G**TAAGTGTCAGAGTGCCCAT CACAACA-3'
positive control (A allele)	5'-TCATAGTTGTTGCAAGCCGAACTGC TGAGAACT**T**TAAGTGTCAGAGTGCCCA TCACAACA-3'

Table 18 – Primers and Probes used for detection of the Breast Cancer Resistance Protein single nucleotide polymorphism C421A.

Statistical analysis

We calculated the distribution of genotypes from the allele frequencies and compared our observations with distributions expected from an assumed Hardy-Weinberg equilibrium using Pearson's X^2 test with two degrees of freedom (df = 2). Odds ratios (OR) and 95 % confidence intervals (CI) were calculated using Fisher's test. A p value < 0.05 was considered statistically significant. All evaluations were performed using Gnu R (http://r-project.org, version 2.8.1) on Microsoft Windows.

Results and Discussion

Determination of genotypes for BCRP C421A, MDR1 C3435T, and MDR1 G2677T

Distributions of all genotypes were according to the Hardy-Weinberg equilibrium. We determined allele frequencies for the individual patient groups (CD, UC) and a combination of both patient groups (IBD). Odds ratios (OR) are given with a 95 % confidence interval (CI). A summary of the results is given in Table 19.

SNP	Group	Genotype frequencies			Allele frequencies		p value (Odds Ratio; 95% Confidence interval)
BCRP C421A		***CC***	***CA***	***AA***	***C***	***A***	
	Controls	15 (0.88)	2 (0.12)	0 (0.00)	0.94	0.06	
	UC	23 (0.68)	10 (0.29)	1 (0.03)	0.82	0.18	0.18 (3.51; 0.63-37.03)
	CD	31 (0.82)	7 (0.18)	0 (0.00)	0.91	0.09	0.71 (1.68; 0.27 – 18.47)
	IBD	54 (0.75)	17 (0.24)	1 (0.01)	0.91	0.09	0.35 (2.48; 0.50 – 24.43)
MDR1 C3435T		***CC***	***CT***	***TT***	***C***	***T***	
	Controls	4 (0.24)	9 (0.53)	4 (0.24)	0.50	0.50	
	UC	5 (0.15)	21 (0.62)	8 (0.24)	0.46	0.54	0.46 (1.76; 0.30 – 9.78)
	CD	7 (0.17)	23 (0.61)	8 (0.21)	0.49	0.51	0.73 (1.36; 0.25 – 6.50)
	IBD	12 (0.17)	44 (0.61)	16 (0.22)	0.49	0.51	0.50 (1.53; 0.31 – 6.20)
MDR1 G2677T		***GG***	***GT***	***TT***	***G***	***T***	
	Controls	12 (0.71)	5 (0.29)	0 (0.00)	0.85	0.15	
	UC	29 (0.93)	5 (0.15)	0 (0.00)	0.93	0.07	0.27 (0.42; 0.08 – 2.20)
	CD	30 (0.79)	7 (0.18)	1 (0.03)	0.88	0.12	0.52 (0.65; 0.15 – 3.04)
	IBD	59 (0.82)	12 (0.17)	1 (0.01)	0.88	0.12	0.33 (0.53; 0.14 – 2.27)

Table 19 – Genotype and allele frequencies of BCRP C421A, MDR1 C3435T, and MDR1 G2677T in UC (n=34), CD (n=38), IBD (n=72), and healthy controls (n=17). All groups are in Hardy-Weinberg equilibrium.

MDR1 C3435T polymorphism

Wild type and mutant alleles were distributed equally in healthy controls. This finding is in accordance with previously published studies [261]. A slight, but statistically insignificantly increased value of 54% in the mutant allele T was only seen in patients with UC (OR: 1.76, CI: 0.30 – 9.78, p< 0.46). Glas et al. [262] and Potocnik et al. [263] have reported similar ratios. No difference was seen for CD (OR: 1.36, CI: 0.25 – 6.50, p < 0.73) or the combined endpoint IBD (OR: 1.53, CI: 0.31 – 6.20, p < 0.5).

MDR1 G2677T polymorphism

Although this SNP is actually tri-allelic, we limited ourselves to the T mutant allele due to our relatively small sample size. Other groups have done the same [256, 263]. At 85%, the wild type allele was more prevalent than the mutant allele in healthy controls. Compared to this, we saw an association with the wild type allele in UC (93%, OR: 0.42, CI: 0.08 – 2.20, p < 0.27) but no difference in either CD (OR: 0.65, CI: 0.15 – 3.04, p < 0.52) or IBD (OR: 0.53, CI: 0.14 – 2.27, p < 0.33).

BCRP C421A polymorphism

The wild type was far more abundant (94%) than the mutant allele. There is a clearly distinguishable trend towards the mutant allele in UC (OR: 3.51, CI: 0.63 – 37.03, p < 0.18), albeit not a significant one. Patients with CD (OR: 1.68, CI: 0.27 – 18.47, p < 0.71) and the combined group of IBD patients (OR: 2.48, CI: 0.50 – 24.43, p< 0.35) showed a similar tendency towards the A allele.

Haplotype analysis

We investigated the associations of combined haplotypes of MDR1 and BCRP SNPs with special attention to the haplotypes previously reported on by other authors [254, 255, 260]. Most strikingly, we found no occurrence of the MDR 1*2 haplotype within any of the subgroups. The strongest association of a BCRP haplotype was for homozygous MDR1 3435CC / BCRP 421CC (X^2: 1.0142, p < 0.30) in UC and MDR1 2677G / BCRP 421A (X^2: 1.5615, p < 0.22), also in UC.

Conclusions

We investigated the prevalence of the SNPs MDR1 C3435T, MDR1 G2677T, and BCRP C421A in blood samples from a Swiss population of healthy volunteers and patients suffering from CD or UC. Genotype and allele frequencies are similar to those found in other European populations. Although the effect of BCRP C421A on IBD is still poorly understood, the mutant allele showed the strongest correlation with UC of all SNPs in this article. An analysis in a larger sample of the population may reveal a statistically significant association.

In our survey of haplotypes, we found no cases of the much discussed MDR1 2677T / G3435. However, our results justify further research into the prevalence of the haplotypes MDR1 3435CC / BCRP 421CC and MDR1 2677G / BCRP 421A, both of which were weakly associated with susceptibility to UC. Again, a larger sample size would be desirable.

The etiology and pathogenesis of IBDs remains poorly understood. The many factors known to influence disease susceptibility and phenotype cannot be pinned down to any single SNP, nor is any haplotype likely to discriminate the many facets of these diseases. [264]

5.1.5. Successful Treatment of a Patient with Crigler-Najjar type II syndrome with St. John's Wort

*Oliver Kummer [1,2], *Felix Hammann [2], Manuel Haschke [2], Stephan Krähenbühl [2]

[1] Department of Gastroenterology and Hepatology, STS-AG, Thun

[2] Divison of Clinical Pharmacology and Toxicology, University Hospital Basel, University of Basel

Correspondence:

Oliver Kummer, MD
Clinical Pharmacology & Toxicology
University Hospital
CH-4031 Basel, Switzerland
Phone: ++41 61 265 25 25
Fax: ++41 61 265 45 60
E-mail: oliver.kummer@unibas.ch

*Oliver Kummer and Felix Hammann contributed equally to the work

Key words: Crigler Najjar, hypericum extract, St. John's wort, hypericum perforatum, hyperforin, induction, icterus, hyperbilirubinaemia, nuclear factor, PXR, pregnane X receptor, constitutive androstane receptor, CAR, peroxisome proliferator-activated receptor, PPARα

Abstract

Introduction: The Crigler-Najjar (CN) syndrome is a very rare disease clinically characterized by unconjugated, nonhaemolytic, severe hyperbilirubinaemia from birth due to an inherited enzyme defect of the uridine 5'-diphospho-glucuronosyltransferase (UGT) isoenzyme 1A1. CN type II syndrome is essentially a cosmetic problem. Possible treatments should therefore not be harmful. However current long-term treatment with phenobarbital is associated with diminished mental activity, lethargy, depression, and teratogenicity. We therefore decided to investigate the effect of hypericum extract, a potent pregnane X receptor mediated enzyme inducer, in a patient with CN type II. **Methods:** The patient was treated with 300 mg St. John's wort dry extract three times daily (Jarsin ® dragées containing the hypericum extract LI160) for seven weeks. Blood samples were obtained between June 21, 2007 and April 24, 2008. To assess enzyme induction, a midazolam kinetic study was performed, before and after hypericum treatment. **Results:** The treatment with hypericum extract was well tolerated by the patient. The area under the plasma concentration time curve (AUC) of midazolam decreased from 4700 ng/mL*min to 2650 ng/mL*min (-44%), indicating cytochrome P450 3A4 (CYP3A4) induction associated with the ingestion of hypericum extract. Mean plasma total bilirubin concentrations were 165±11 µmol/l (mean ±SD, n = 7) without any medication and 112± 8 µmol/l (mean ±SD, n = 6) during treatment with hypericum extract. Compared to the plasma concentration without any treatment, hypericum extract was associated with an average 32% decrease in total plasma bilirubin. The highest measured total bilirubin plasma concentration in the study period without treatment was 204 µmol/L, whereas the lowest concentration with hypericum extract treatment was 93 µmol/L. **Conclusion:** Treatment with hypericum extract in a CN type II patient is well tolerated and showed a comparable reduction of the bilirubin plasma concentration as the current, more toxic treatment with phenobarbital.

Introduction

Glucuronidation represents a major pathway in mammalians for elimination of lipophilic compounds [265]. The transfer of glucuronic acid is catalyzed by members of the uridine 5'-diphospho-glucuronosyltransferase (UGT) family [266, 267]. Each UGT isoenzyme exhibits a profile of substrate and tissue specificity, which is overlapping with other UGT isoenzymes [34].

The Crigler-Najjar (CN) type II syndrome is clinically characterized by unconjugated, nonhaemolytic severe hyperbilirubinaemia from birth [268, 269], due to an inherited enzyme defect of the UGT isoenzyme 1A1. UGT1A1 is the only enzyme that contributes significantly to bilirubin glucuronidation which is necessary for biliary excretion [270]. CN type II patients have, compared to patients with CN type I, only a partial UGT1A1 deficiency with serum bilirubin levels between 60 and 340 µmol/l [271-274]. Treatment of CN type II patients with 5 to 10mg phenobarbital per kilogram body weight leads to a relevant reduction of serum bilirubin levels within two weeks after treatment start [273-278]. The expression of the UGT1A1 has been shown to be induced by the activation of constitutive androstane receptor (e.g. phenobarbital [274,

279]), peroxisome proliferator-activated receptor α (e.g. clofibrate [280]) and furthermore pregnane X receptor (e.g. dexamethasone [280], progesterone [281, 282]). Since CN type II syndrome is essentially a cosmetic disease, treatment should not be harmful. Despite the rapid tolerance to the sedative effect of phenobarbital within three to four days, diminished mental activity, lethargy, depression, and teratogenicity are limiting factors of long-term treatment with phenobarbital [283, 284]. We therefore decided to study the effect of hypericum in a patient with CN type II. Hypericum (St. John's wort) is an effective and well tolerated medication for mild depression [285]. Hypericum extract LI160 is a potent PXR activator with a high hyperforin content, the main inducing component of hypericum [286-288].

Materials and Methods

Patient

A non smoking, 24 year old Caucasian woman with jaundice from birth was enrolled for the study. The clinical suspicion of a CN type II syndrome could be confirmed by the identification of point mutations in the nucleotide positions 211 and 1456 of the UGT1A gene. These mutations have been described in two patients with phenobarbital-sensitive CN type II syndrome [289, 290]. Because of dysmenorrhea and mild acne vulgaris, the patient was treated with ethinylestradiol 0.03 mg and dienogest 2 mg since June 2006. The patient ingested no other drugs and had no other diseases.

Study design

We performed a single-case, open-label, three-period enzyme induction study in this CN type II patient. Since CN type II is a very rare disease, we could not find other patients for inclusion into the study. The primary study endpoint was changes in the serum bilirubin associated with the ingestion of hypericum. Because of the known high intra-individual variability of bilirubin levels, we tried to avoid known influencing factors as well as possible. The patient was told to refrain from sunbathing, tanning bed sessions, weight reducing diets, sleep deprivation, vitamin and herbal supplements, as well as grapefruit products during the whole investigational period from May 2007 till end April 2008. A midazolam (MDZ) kinetic study was performed to asses enzyme induction by the hypericum extract [288, 291-293]. The study was conducted in conformity with the principles of the Declaration of Helsinki and approved by the Ethics Committee of the State of Basel, Switzerland. The patient gave written informed consent before undergoing any study-related procedures.

Blood sampling for bilirubin determinations

A total of 19 blood samples were obtained between June 21, 2007 and April 24, 2008. Six blood samples were obtained under oral contraception with ethinylestradiol 0.03 mg and dienogest 2 mg between June 21, 2007 and October 22, 2007. End of October 2007, oral contraception was interrupted. After a washout phase of seven days another five blood samples were obtained without any treatment between November 6, 2007 and January 8, 2008. On February 8th, the patient started the treatment with 300mg hypericum

extract three times daily (Jarsin® dragées LI160, Vifor SA, Villars-sur-Glâne, Switzerland). Thereafter, between February 12 and March 20, 2008, six blood samples were obtained while the patient was treated with hypericum extract. On March 20, 2008, we stopped treatment with hypericum extract. After a washout phase of one month, we obtained two additional blood samples (April 17 and April 24, 2008) without any treatment.

Midazolam pharmacokinetics

We verified PXR activation and subsequent CYP3A4 induction by assessing the pharmacokinetics of a 2mg intravenous of MDZ (Dormicum®, Hoffmann-LaRoche Ltd., Basel, Switzerland) immediately before starting treatment with hypericum extract (February 8, 2008), and after treatment with hypericum extract for three weeks (February 28, 2008). Venous blood samples were obtained immediately prior to and 5, 20, 60, 120, 240, 360, 480, 600 min after MDZ injection using Monovette tubes (Sarstedt, Nürtingen, Germany), containing lithium heparin as an anticoagulant. Samples were put on ice and centrifuged at 3'000 g over 5 min at 4° C within 30 minutes after collection. The resulting plasma was stored at -70°C until MDZ analysis. MDZ and its metabolites were analyzed by liquid chromatography–mass spectrometry/mass spectrometry as described previously [294]. Free concentrations of MDZ, 1-OH-hydroxymidazolam (1′-OHMDZ), and 4-OH-hydroxymidazolam (4-OHMDZ) were determined after ultra filtration of the plasma samples. The metabolite 1′-OHMDZ glucuronide was measured as 1′-OHMDZ after deglucuronidation using the same liquid chromatography–mass spectrometry/mass spectrometry method.

Pharmacokinetic analysis

Plasma MDZ, 1′-OHMDZ, 4-OHMDZ and glucuronide data were analyzed using compartmental (MDZ) and non-compartmental methods (1′-OHMDZ, 4-OHMDZ) (WinNonlin, version 5.01, Pharsight Corp., Mountain View, CA, USA).

Results

Bilirubin plasma levels

Altogether, 19 blood samples were obtained between June 21, 2007 and April 24, 2008. Mean plasma total bilirubin concentrations were 165 ±11 µmol/l (mean ±SD, n =7) without any medication, 134 ±8 µmol/l (mean ±SD, n=6) under treatment with ethinylestradiol/dienogest, and 112 ±8 µmol/l (mean ±SD, n =6) during treatment with hypericum extract (Figure 54). Compared to the plasma concentration without any treatment, hypericum extract was associated with an average 32% decrease and ethinyl/dienogest with an average 19% decrease in total plasma bilirubin.

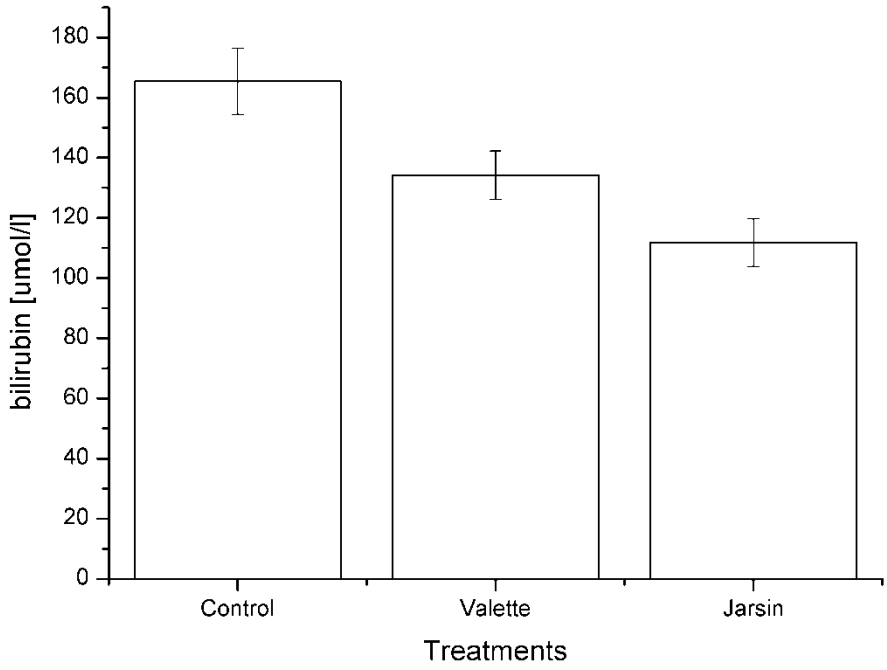


Figure 54 – Bilirubin plasma levels under treatment with the oral contraceptive Valette ®, the anti-depressant Jarsin ®, and control without medication.

Pharmacokinetics of midazolam

After induction with hypericum extract the area under the plasma concentration time curve (AUC) of midazolam decreased from 4700 ng/mL*min to 2650 ng/mL*min (-44%). AUC of 1′-OHMDZ before (597 ng/mL*min) and after induction (566 ng/mL*min) didn't changed relevant, whereas a marked decrease of the 4-OHMDZ before (121 ng/mL*min) and after induction (48 ng/mL*min) was found. Results are summarized in Table 20.

Tolerability of hypericum extract

The treatment with hypericum extract was well tolerated, except a feeling of slight facial skin prickling without visible skin changing's for some minutes after a one hour lasting walking tour on a sunny day.

Midazolam bolus 2mg i.v	before hypericum treatment	After 21 d hypericum (900 mg/d)
AUC (0,∞) (ng ml^{-1} min)	4700	2647
$t_{1/2\alpha}$ (min)	14	6
$t_{1/2\beta}$ (min)	133	128
C_{max} (ng ml^{-1})	49	84
1'-hydroxymidazolam		
AUC (0,∞) (ng ml^{-1} min)	597	566
4'-hydroxymidazolam		
AUC (0,∞) (ng ml^{-1} min)	121	48

Table 20 – Pharmacokinetic evaluation of midazolam in plasma before and after treatment with hypericum extract.

Discussion

MDZ is rapidly metabolized by CYP3A4 to its main metabolite 1′-OHMDZ. The induction of this step have been shown in our patient, the MDZ AUC dropped by 44% after induction with hypericum. The following glucuronidation is performed mainly by UGT2B7 and UGT2B4 (O-Glucuronidation) and to minor extend by UGT1A4 (N-Glucuronidation) [295, 296]. Because of induction of the glucuronidation we expected a relevant decrease in the 1′-OHMDZ AUC under treatment with hypericum, but AUC of 1′-OHMDZ before and after induction was similar. This may be caused by the additional point mutations in the nucleotide positions of 1456 in the codon 5 of the UGT1A gene encoding for UGT1A4, performing N-glucuronidation of 1'-hydroxymidazolam.

Treatment with hypericum extract in a CN type II patient is well tolerated and showed a comparable reduction of the bilirubin plasma concentration as the current, more toxic treatment with phenobarbital. A case series of five Crigler Najjar type II patients compared the highest measured serum bilirubin concentrations before treatment with phenobarbital with the lowest measured concentrations after induction treatment. The highest measured serum bilirubin concentrations fell by between 33 and 77% compared with the lowest measured bilirubin concentration after treatment with phenobarbital [275]. If we compare the highest bilirubin concentrations in our patients without treatment with the lowest concentration with hypericum extract treatment we find a maximal reduction of the bilirubin plasma concentration of 54%.

After induction with hypericum, the MDZ AUC0-12h decreased by 44% compared with basal values. From literature we know that the extract LI 160 (Jarsin®) decreased MDZ AUC0-12h by 79.4% (95% CI 88.6; 70.1) [35]. The increased MDZ elimination is only a surrogate marker of induction and demonstrates that our patient was compliant to the study medication.

Although CN type II patients become tolerant to the sedative effect of phenobarbital within three to four days [273], diminished mental activity, lethargy, depression, and teratogenicity are limiting factors of long-term treatment with phenobarbital [283, 284]. On the other hand, the risk of gallbladder disease, pancreatitis, and malignancies are adverse reactions, limiting the use of clofibrate [297, 298]. In contrast hypericum extract was well tolerated by our patient, as expected from patients with depressions treated with hypericum extract [299]. The most common adverse effects of hypericum extract are gastrointestinal symptoms, dizziness and tiredness. The often discussed phototoxic skin reaction were described especially in studies where patients were treated with pure hypericin [300-302] whereas photosensitivity is extremely rare in patients treated with hypericum extract [303, 304].

5.1.6. Pulsatile transdermal delivery of nicotine to male smokers

Felix Hammann [1*], Oliver Kummer [1*], Georg Imanidis [2], Juergen Drewe [1]

[1] Department of Clinical Pharmacology and Toxicology, University Hospital Basel, University of Basel

[2] University of Applied Sciences Northwestern Switzerland, Dept. of Pharmaceutical Technology

[*] both authors contributed equally to this work

Submitted to: Journal of Controlled Release

Address for Correspondence:
Prof. Juergen Drewe
Department of Gastroenterology
University Hospital of Basel
Petersgraben 4
CH-4031 Basel Switzerland
Email: juergen.drewe@unibas.ch

Abstract

Smoking is a major risk factor for many preventable diseases and presents a great threat to the public health. Nicotine substitution is still one of the mainstay therapies, although nicotine patches and gum do not give smokers the same pharmacokinetic profile as cigarette consumption and may therefore be poorly effective. Here, we present a computer-controlled delivery system for pulsatile transdermal administration of nicotine. We enrolled twelve male volunteers who consume more than 20 cigarettes/day and established tolerability with three increasing doses of nicotine (6.7 mg/d, 13.4 mg/d, and 26.7 mg/d). Furthermore, we showed the feasibility of controlled release administration of nicotine in dose linear manner with dose application at precise intervals and amounts which yield effective plasma peaks. After additional efforts have been made to further miniaturize the system, it could find use not only in nicotine replacement therapy, but also for other indications such as patient-controlled analgesia.

Introduction

Tobacco use is one of the major risk factors for cardiovascular and pulmonary morbidity such as coronary heart disease, peripheral artery disease, chronic obstructive pulmonary disease (COPD), lung cancer, and infections [305-308]. Smoking also increases the risk of cancer of other organs [309]. Half of the estimated 1-2 billion smokers worldwide will die from these and other smoking-related illnesses [310] – not accounting for the negative effects of passive exposure to smoking. Quitting smoking at any age is beneficial as duration and amount of tobacco use correlates with disease progression and severity [311]. Effective smoking cessation strategies are therefore of paramount importance to the public health.

Mechanisms of tobacco addiction

Nicotine ((S)-3-(1-methyl-2-pyrrolidinyl) pyridine) is the main addictive agent in tobacco products. When smoked, nicotine is distilled and rapidly taken up into the blood from which it readily penetrates the blood-brain barrier. This is shown by high, short concentration peaks 10-20 s after inhalation [312]. Nicotine has a high abuse liability, leading to physical dependence and compulsive abuse [313]. The addictive potential of nicotine is rated as high as that of cocaine or heroin, but without the behavioral disruption that goes along with other substances of abuse [314]. Development of nicotine addiction may be modulated by the rate of nicotine uptake, with higher invasion rates into the CNS showing a higher risk [311].

Therapeutic options

After willingness to quit has been assessed, treatment of nicotine dependence usually involves continued psychological counseling and pharmacological support. First-line drug therapies are nicotine replacements (e.g. patches, gums, or sublingual forms) and non-nicotine agents such as the partial nicotinic receptor agonist varenicline [315]. Second-line agents, e.g. antidepressants or clonidine [316], help in treating withdrawal symptoms. The overall performance of nicotine replacement is, however, still disappointing,

with only about 15% of smokers seeking treatment actually giving up the habit [317], This is probably because available forms of nicotine replacement therapies (NRTs) do not give the patient the immediate high and short peaks as caused by smoking [318].

Pulsatile delivery systems

In pulsatile delivery, an interval of no release is followed by controlled, quick, and complete release of a drug dose. Site-specific pulsatile release is desirable in drugs with a high first-pass effect that target distant parts of the intestine. Time-controlled systems, such as the one presented in this study, release at set time points to produce a specific chronopharmacological profile or on demand [319, 320].

Study rationale

In this study, we evaluated the feasibility of pulsatile transdermal delivery of nicotine to mimick the nicotine plasma peaks in inhalative smoking with a new microprocessor controlled system that dispenses the substance from a drug reservoir. Furthermore, we aimed at establishing the tolerability of nicotine with pulsatile release kinetics in otherwise healthy male smokers.

Materials and methods

Volunteers

Twelve male heavy smokers, aged 28.3 (20.7-47.2) years (mean (range)), were enrolled for the study. Weights were within normal limits (body mass index: 23.1 (20.1-25.1) kg/m^2 (mean (range)) and heights were 178 (170-184) cm (mean (range)). Prior to participation, subjects had to give written informed consent. Good health was assessed with a medical history, physical examination, and laboratory controls of blood and urine. Tobacco use of 20 cigarettes/day or more was assessed with a urine cotinine test. None of the volunteers was taking medication or suffered from an allergic condition. Two additional volunteers were included in the study but dropped out because of personal reasons.

Study design

The study was a single-center, open-label, three-periods, dose-escalating study, and was approved by the local state ethics committee *(Ethische Kommission beider Basel)*. We consecutively assigned subjects to three treatment groups, where the dosing device dispensed nicotine at rates of 0.67 mg/h, 1.34 mg/h, and 2.67 mg/h in two peaks separated by 8 h. These rates correspond to total doses of 6.7 mg, 13.4 mg, and 26.7 mg, respectively. The rationale for these doses was based on the nicotine plasma levels in regular smokers [321] and the release rate required was determined in a pilot experiment.

Subjects arrived at 6:00 a.m. of each test day at the Clinical Research Center (CRC) of the University Hospital Basel, after at least 12 hours fasting and abstaining from tobacco. Vital signs of the subjects,

adverse events, and any new medical conditions which might have arisen were recorded. For nicotine administration, the dosing device was placed on the flector side of the leading lower arm and a venous catheter was inserted into the cubital vein on the opposite side.

Vital signs were recorded and blood was drawn pre-dose and then at hourly intervals for 16 hours and 24 h after the start of administrations. The device remained fixed on the arm for 16 hours, releasing nicotine at the assigned dispensing rates reported for 2 min at 1h and 8h, after which it was removed and subjects remained on the ward for monitoring until they were discharged in the morning.

The three consecutive treatments were separated by a wash-out period of at least three days. Adverse events were continually recorded and subjects were admitted to the next period only if the preceding dose had been tolerated well.

Analysis and statistical evaluation

Whole blood was drawn and immediately centrifuged (3'000 g over 5 min at 4° C). Supernatant plasma was stored at -20° C. For processing, samples were thawed at room temperature and vortexed. Then, 400 µl of plasma were added to 400 µl of 20% w/v trichloroacetic acid, and, after further vortexing, kept at -17° C for 30 minutes. Subsequently, the samples were centrifuged at 14'000 g for 20 minutes. The supernatant was collected and used directly for HPLC-MS analysis.

Agilent equipment was used, consisting of a G1312A binary pump, a G1379B degasser, a G1367B high performance thermostated autosampler, a G1314B UV detector, and a 6130 single quadrupole MS detector with an atmospheric pressure electrospray ionization source. Nicotine was quantified on the MS signal using positive mode selected ion monitoring at m/z 163.1, a capillary voltage of 4000 V, drying gas flow 10 l/min, nebulizer pressure 30 psig, drying gas temperature 350°C, a fragmenter setting of 70 V and gain setting 3. Chromatography was performed on a 125/2 mm Nucleosil 100-5 C8 ec column (Macherey-Nagel) with an injection volume of 100 µl. A mobile phase consisting of 7.5% methanol and 92.5% water and containing 1% acetic acid (96%) and 20 mM ammonium acetate was used in isocratic mode. Quantification was performed against a set of external standard solutions of nicotine that were prepared in water and treated in the same way as the plasma samples.

With this method, a linear peak area response was obtained at concentrations of injected sample between 1 and 40 ng/ml. Over this concentration range, recovery of the drug from plasma determined by spiking blank human plasma with nicotine was found to be between 85 and 110%. The limit of quantification was 1 ng/ml with a variation between 5 and 10%, and the limit of detection was approximately 0.3 ng/ml.

The following pharmacokinetic parameters were determined: area under the curve (AUC) was calculated with the linear trapezoidal rule, C_{max}, and T_{max} of the first (1-7 h) and second peak (8-16h) were determined by inspection of the plasma concentration/time data. Statistical analysis was performed in SPSS for Windows software (version 15.0; SPSS Inc., Chicago Ill). The level of significance was $p \leq 0.05$.

Materials

Chrono Therapeutics Inc. has developed a watch-sized device (ChronoDose™, seen in Figure 55) to deliver drugs transdermally by microprocessor controlled delivery. For this study, we used prototypes of the device, which consisted of a substance reservoir and a replaceable membrane of 10 cm^2. The devices were connected to notebook computers via a parallel connection that controlled drug release and a USB cable for power.

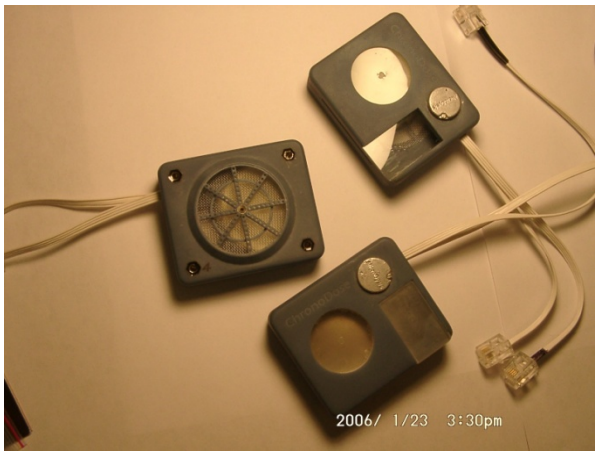


Figure 55 – Prototypes of the ChronoDose™ transdermal drug delivery device (picture courtesy of Chrono Therapeutics, Inc.)

Nicotine drug substance was obtained in GMP-quality for human use from Siegfried Ltd., Zofingen, Switzerland. Solutions and transdermal systems were prepared at the University of Applied Sciences Northwestern Switzerland under the supervision of Prof. G. Imanidis

Results and Discussion

Pharmacokinetic parameters

Even for small lipophilic compounds such as nicotine, the efficiency of transdermal delivery is influenced by a large number of parameters (skin thickness, sweat production, etc.). Because of the expected interindividual variability and lack of experience in human subjects for this device, we used an initial flow rate that was estimated to yield levels just above the level of quantification. This was observed indeed after administration of the lowest release rate. Plasma levels were only quantifiable for 7 out of 12 volunteers and the dose was well tolerated. The results of the nicotine plasma analysis are summarized in Table 21 and in Figure 56.

Flow rate ($mg/cm^2/h$)	AUC (ng*h/ml) 0-24 h	C_{max} (ng/ml) 0-7h	T_{max} (h) 0-7h	C_{max} (ng/ml) 8-16h	T_{max} (h) 8-16h	n
0.67	56.7 ± 10.0	4.9 ± 1.5	3.7 ± 0.6	5.7 ± 1.1	10.9 ± 0.9	7
1.34	131.4 ± 19.4	7.7 ± 1.4	4.9 ± 0.5	13.0 ± 1.9	10.2 ± 0.6	11
2.67	262.8 ± 39.7	18.2 ± 2.8	4.8 ± 0.4	24.7 ± 4.1	10.0 ± 0.4	12

Table 21 – Pharmacokinetics of two increasing transdermal doses of nicotine in healthy male smokers (AUC, area under the concentration-time curve; Cmax, peak plasma concentration; Tmax, time of peak plasma concentration; n, number of samples. Data represented as means +/- SEM.)

Doses (6.7 mg, 13.4 mg, and 26.7 mg) were administered at 0h and 8h after start of administration and the respective peaks (T_{max}) were seen on average (± SEM) at 4.5 ± 0.4 h and 10.4 ± 0.4h, respectively. AUC increased in linear with the dose ($p < 0.02$, $R^2 = 0.998$) and so did C_{max} for the second peak ($p < 0.04$, $R^2=0.994$) while we only observed borderline significance for the linear increase of C_{max} of the first peak ($p = 0.086$, $R^2 = 0.964$).

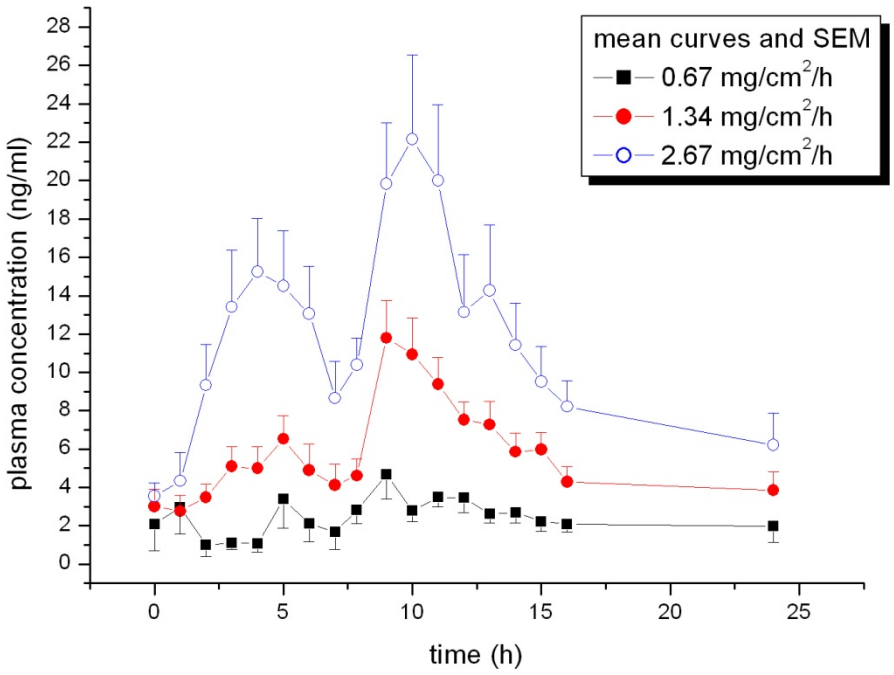


Figure 56 – Plasma concentration of nicotine in response to three increasing pulsatile transdermal doses.

Adverse events

At the lowest dose, several volunteers expressed a craving for nicotine. Mild to moderate erythema was seen at the site of membrane placement in all volunteers at all doses but had disappeared completely within 1-2 days after removal of the patch without need for treatment. On the next study day, no signs of erythema were visible.

Discussion

All three doses were well tolerated except for local irritation of the skin, a common adverse effect of topical nicotine [322]. Within the variability commonly seen in transdermal systems, the prototypes of the ChronoDose™ device used in this study were capable of delivering nicotine at precise intervals and amounts with dose linearity.

Conclusion

In this work, we present a computer-controlled system for transdermal delivery of nicotine and show its tolerability and potential benefits in NRT. Apart from NRT, this system may also find application in the management of postoperative pain [323, 324] or, more generally, as dispenser of other substances such as analgetic agents (patient-controlled analgesia) or isosorbide dinitrate (ISDN) [325]. The ChronoDose™ device prototype we used here requires an external computer to control it, but an embedded microprocessor design is being developed.

Conflict of interest

This study was supported by Chrono Therapeutics, Inc., Hamilton, NJ, USA.

5.2. Isolated project

5.2.1. An Automated General Unknown Screening for Drugs and Toxic Compounds in Human Serum Using Liquid Chromatography-Tandem Mass Spectrometry

Stefan Sturm [1], Felix Hammann [2], Juergen Drewe [2], Hans H. Maurer [3] and André Scholer [1]*

[1] University Hospital of Basel, Laboratory Medicine, Clinical Chemistry Laboratory, Basel, Switzerland

[2] University Hospital of Basel, Clinical Pharmacology and Toxicology, Basel, Switzerland

[3] Saarland University, Department of Experimental and Clinical Toxicology, Institute of Experimental and Clinical Pharmacology and Toxicology, Homburg/Saar, Germany

*Corresponding Author.
André Scholer
Laboratory Medicine
Clinical Chemistry Laboratory
University Hospital of Basel
Petersgraben 4
4031 Basel
Switzerland
Telephone number: 0041-61-265 42 36
Fax number: 0041-61-265 46 00
E-mail: scholera@uhbs.ch

Keywords: general unknown screening; data-dependent acquisition; toxic compounds

Abstract

A fully automated general unknown screening using liquid chromatography-mass spectrometric method applying data-dependent acquisition was developed to identify toxicologically relevant substances in serum and urine. The method was developed to get an alternative to the HPLC Remedi System from Biorad the support of which will be stopped at the end of 2008. A library including more than 405 specters of about 365 compounds (main drugs and important metabolites) was established. An easy to use program was created to automate and accelerate library search. Drugs were identified based on their relative retention times, molecular ions and fragment ions. Limits of detection of the majority of 100 out of 350 compounds (67%) were lower than 100 μg/l. The developed LC-MS-MS system seems to be a valuable alternative to other general unknown screening methods allowing fast and specific identification of drugs in serum and urine samples.

Introduction

A non-target screening for detection of drugs and toxic compounds is called general unknown screening (GUS) procedure or systematic toxicological analysis (STA). This procedure is an analytical method designed to detect and identify xenobiotics in biological fluids, which is necessary for confirmation of the diagnosis of an acute poisoning with drugs or other exogenous compounds. Rapid and comprehensive screening procedures are therefore necessary.

For current STA procedures in clinical and forensic toxicology, automated immunoassays for the most common drugs of abuse in combination with chromatographic techniques coupled to specific detectors are often used. Gas chromatography-mass spectrometry is so far the gold standard for this purpose [326-328]. Big advantages of this method are the content of compounds in the library (> 7000) and the transferability of the library to GCMS systems of different manufacturers.

Its application however is limited to non-polar, volatile and thermally stable compounds. In addition, derivatization is necessary for detection of polar compounds such as metabolites, which complicates the screening procedure, but allow to detect compounds with different pKa values to be analyzed in one GC run. HPLC coupled to UV diode array detection (DAD) overcomes these limitations [329-332]. However it shows reduced separation efficiency and the detection of compounds is not as specific and reliable as compared to GC-MS [326].

Therefore, in the last years, the combination of mass spectrometry with liquid chromatography has been evaluated for screening analysis. It was shown to be very sensitive, precise, specific, universal and very fast if coupled to an automated extraction system [333-340]. Several authors described screening methods with LC-MS that apply in-source collision induced dissociation (CID), LC-MS-MS in the multiple reaction monitoring mode, and LC-MS-MS using data-dependent acquisition (DDA) [333-342].

With single MS, the mass spectrometer operates in the scan mode and applies in-source CID. The sample is screened at variable orifice voltages [337, 339-341]. Reconstructed spectra can be obtained and compared with in-source CID spectra libraries. MS-MS data have the advantage of providing a higher specificity and selectivity and more structural information than single MS. This mode has been shown to be helpful when an unknown substance has to be identified. Although LC-MS-MS in the multiple-reaction monitoring mode can be applied to a high number of previously selected compounds, this number is limited [336].

LC-MS-MS using DDA seems to be the best procedure for simultaneous screening and identification of unknown compounds using a single chromatographic run. In this procedure, ions that exceed a pre-set threshold are fragmented by CID and the resulting fragments are measured in the product-ion scan mode [333-335]. This technique is highly specific and selective. Spectra results from a single ion and the origin of the spectra are registered. This study presents an approach of a fully automated screening procedure with a specially created library search program to perform compound identification. The chosen procedure consisted of on-line solid-phase extraction (SPE) and LC-MS-MS using DDA.

Experimental

Materials

Test substances obtained from various pharmaceutical companies were of pharmaceutical purity. Organic solvents and reagents were of analytical grade. Acetonitrile and methanol were purchased from Merck (Darmstadt, Germany), ammonium formate from Aldrich (Steinheim, Germany) and formic acid from Fluka (Buchs, Switzerland). De-ionized water was generated with a Milli-Q water purification system from Millipore (Kloten, Switzerland). Drug free serum was purchased from Biorad (Reinach, Switzerland). Serum and urine samples for the comparison study were made anonymous after performing the analysis of routine requests and used without any personal information for the comparison between different methods (with the verbal agreement of the Ethic Commission of the two cantons of Basel-Landschaft and Basel-Stadt).

Apparatus

The chromatographic system consisted of a Rheos 2000 Micro HPLC pump from ThermoFinnigan (Allschwil, Switzerland) and a Midas Symbiosis Autosampler from Spark Holland (Emmen, Netherlands) using a 100 μL loop. The detector was a ThermoFinnigan LCQ Advantage MAX ion trap mass spectrometer equipped with an atmospheric pressure chemical ionization (APCI) device and the Xcalibur software. Automated solid-phase extraction was performed using a Prospekt 2^{TM} from Spark Holland consisting of an automatic cartridge-exchange module, dual cartridge clamps and a solvent delivery unit. HySphere Resin GP cartridges were purchased from Spark Holland.

Methods

Standard solutions

Separate stock solutions were prepared in methanol-water (1:1, v/v) at a concentration of 100 mg/l. Serum standards were prepared by spiking with stock solutions of drug mixtures to get concentrations ranging from 0.005 to 4 mg/l, resulting in a set of standards with the following concentrations: 0.005, 0.010, 0.025, 0.050, 0.100, 0.250, 0.500, 1.000, 2.000 and 4.000 mg/l. d_3-Benzoylecgonine was prepared as internal standard (IS) at a concentration of 5 mg/l.

The results of the new method were compared to analysis performed on a Remedi Systems from Biorad (Remedi HS and Remedi Benzodiazepine [343]), to a in house developed LC-MS-method (Finnigan Navigator) following the procedure described by Bogusz et al. and to a full-scan GC-MS screening method [344] applied in the Department of Experimental and Clinical Toxicology, Saarland University, Homburg/Germany.

Extraction procedure

One ml of serum or urine was acidified by addition of 20 μL concentrated formic acid (cleaving a possible protein binding of drugs) and 100 μL of the IS solution were pipetted into each sample. On-line SPE and elution were performed using the Prospekt 2 system. The HySphere Resin GP (Spark Holland) cartridge was conditioned with 1 ml of methanol (5 ml/min) and with 1 ml of water (5 ml/min). A 100-μl-aliquot of the serum was loaded on the cartridge. The sorbent was washed with 1 ml of water (2 ml/min), and eluted directly with the mobile phase over 15 min. In this study, cartridges were used only once in order to avoid possible contamination by proteins which could reduce the extraction rate and destroy the separation column.

Evaluation of matrix effects and process efficiency

Possible influences by matrix effects were studied with three different methods. In the first test, process efficiency was determined. For calculation of the process efficiency expressed in percent (PE%), the peak area ratio (i.e. the peak area of the drug of interest was divided by the peak area of the IS) obtained after the on-line extraction of a serum sample and compared to the peak area ratio obtained after direct injection of the same amount of an aqueous solution into the LC-MS-MS system. (The PE% of the IS (84%) was also considered in the calculations (PE% = (Peak area ratio of a serum sample spiked before extraction / peak area ratio of an aqueous solution x 100).

The second procedure is based on the post-column infusion of an analyte in a chromatographic run of blank serum or urine (urine samples were tested when the method was in routine). The signal was compared to the signal obtained with post-column infusion of the parameter to be tested into the eluent of the corresponding blank matrix extract. In the last experiment, blank samples (serum and urine) used as negative controls were analyzed.

Liquid chromatography

The chromatographic separation was performed on a CC Nucleodur C18 Gravity 3 μm column (4 x 125 mm) with an integrated guard column 3 μm (4 x 8 mm) from Macherey-Nagel (Oensingen, Switzerland). The mobile phase was delivered at a flow rate of 400 μL/min. Each chromatographic run was performed with a binary, linear A/B gradient (Solvent A was 10 mmol/l ammonium formate, pH 3.0. Solvent B was 90% acetonitrile, 10% 10 mmol/l ammonium formate, pH 3.0.). The program was as follows: 0-1 min, 6% B; 1-8 min, 6 to 100% B; 8-20 min 100% B; 20-23 min column equilibration with 6% B.

Mass spectrometry

The following APCI inlet conditions were applied. The heated vaporizer was kept at 465 °C. Both the sheath gas set at 60 relative units and the auxiliary gas set at 15 relative units were nitrogen. The capillary entrance to the ion trap was at an offset of 28 V in the positive mode, -4 V in the negative mode and was maintained at 220° C. The corona current was 5 μA. Table 22 shows the data dependent and global data dependent settings.

Data dependent settings	
Default charge state	1
Default isolation width (m/z)	4.0
Normalized collision energy (% pos/neg))	40.0 resp. 35.0
Minimal signal required	20’000

Global data dependent settings	
Exclusion mass width (m/z)	0.5
Reject mass width (m/z)	1.0
Dynamic exclusion	enabled
Repeat count	1
Repeat duration (min)	0.5
Exclusion list size	25
Exclusion duration (min)	0.5
Exclusion mass width (m/z)	0.5

Table 22 – Data dependent and global data dependent settings

DDA was used, generating a full-scan between 80 and 750 atomic mass units in the first mode. If ions exceeded the preset threshold, a MS-MS spectrum of the most intense ion of the previous full-scan was acquired in the second mode. The maximum injection time was set to 50 ms, and three micro-scans were collected for each data point. Normalized collision energy was 40.0% in the positive mode and 35.0% in the negative mode. Dynamic exclusion was enabled meaning that a refractory period was applied to the last selected ion. The refractory period was thirty seconds.

Evaluation of the limit of detection (LOD)

For qualitative purposes only the LOD of each substance in the library of a GUS method is important in order to know the specific performance of the procedure. The LOD's were measured for about 100 substances out of 350 (all other LOD's will be measured step by step together with routine analysis) by analyzing each substance after spiking to drug free serum with decreasing concentrations in the range of 4.000 mg/l to 0.005 mg/l (two times the same procedure and defining the limit by the mean of the two results).

Mass spectral library

Standard solutions were prepared in methanol-water (1:1, v/v) at a concentration of 1-2 mg/l. Two mass spectral libraries were created, one for each ionization mode (positive and negative), by injecting 20 µl of these solutions directly without HPLC separation into the MS system. The obtained MS-MS spectra were added to the library. Relative retention times (RRT) were acquired by actual LC-MS (-MS) analysis running a mixture of each compound spiked to serum (and in a new trial in urine) with the IS (see also under results of reproducibility, chapter 3). The RRT's have a 1 to 2% reproducibility. MS-MS data obtained from a chromatographic run were compared to the MS-MS library using the NIST Mass Spectral Program 2.0 from Thermo Finnigan. A computer program (XcLibraryScreening) was created to automate the searching process and to include the RRT and the molecular ion in the identification of unknown compounds.

Mass spectral library search program (XcLibraryScreening)

To automate the search process and to combine the MS-MS library with the RRT of each substance, we developed a Microsoft Windows application (XcLibraryScreening). It was written in Microsoft Visual Basic .NET and requires the Microsoft .NET Framework as well as a running copy of the NIST Mass Spectral Program 2.0 from ThermoFinnigan with a correctly configured MS-MS library. RRTs are stored in a separate comma separated value (CSV) file which can be edited in any spreadsheet application or text editor. The user can configure searches with XcLibraryScreening by giving the RT of the internal standard ("RT IS" in the setup dialog), mass-to-charge ratios ("m/c +/-"), and the match factors ("Match Factor" and "reverse Match Factor"). Output is presented in a separate window (Figure 59) and can be saved to a text file for reference.

Results and discussion

On-line SPE was chosen as an extraction technique because this procedure is universal, rapid and can be automated. Therefore, this method is becoming popular in bio-analytical analysis [334]. The Prospekt SPE can be linked to the LC-MS-MS instrument [333]. This system couples and automates sample extraction and instrumental analysis. Benefits of this technology include improved precision of all extraction steps. The method has a time saving advantage compared to other techniques because the evaporation of the liquid sample extract is not necessary. In addition, the procedure presented in this study extracts acidic, neutral as well as basic drugs. A polymer-based sorbent was chosen. This sorbent is stable over a large pH range and does not have interfering secondary groups (unbounded hydroxyl silanols) compared to classical hydrophobic phase.

The extraction is an on-line procedure and the elution solvent is the mobile phase. In addition the mobile phase consists of a gradient. The PE% of 10 drugs from different substance groups was determined to have an idea of the extraction recovery at a concentration of 1 mg/l for each substance (Table 23). Among the ten substances, the PE% was between 80% and 119%. Due to this high process efficiency of the ten tested substances it is supposed that a high extraction rate was achieved and that our method seems not to be considerably affected by suppression or enhancement of ionization due to sample matrix. ME = peak area in sample spiked after extraction / peak area of neat standard

The post-column infusion of the model substances codeine and benzoylecgonine in a chromatographic run of blank serum and urine compared to eluent only indicated that in general no change in the ionization process of the two tested substances were found due to co-eluting compounds (data not shown).

Drug	Process efficiency	Drug	Process efficiency
Morphine	85%	Torasemide	80%
Olanzapine	119%	Propranolol	95%
Ephedrine	94%	Bupivacaine	105%
Gliclazide	91%	Phenprocoumon	97%
Citalopram	106%	Phenolphtaleine	118%

Table 23 – Process efficiency (%) of 10 substances from different substance groups. Each substance in a concentration of 1 mg/l in a drug free serum

The separation of the drugs was carried out under acidic conditions (pH = 3) in order to limit secondary interaction on the free silanol groups of the C_{18} separation column. The first peak eluted at 5.9 min (morphine) and the last at 18.4 min (delta-8-THC). The absolute retention times (RT) as an example for 100 out of 365 compounds are shown in Table 24 (all retention times of the 365 compounds are summarized in an excel table in the library together with all needed information for the search program). Analysis of eight different plasma samples was performed on the same day and on different days to study

the intra- and inter-assay precision of the IS retention times. The intra- and inter-assay precision of the IS were 6.92 ± 0.02 minutes and 6.91±0.03 minutes respectively.

Drug	Retention time (min)	Drug	Retention time (min)
Morphine	5.89	Imipramine	8.40
Amiloride	6.10	Phenobarbital	8.45
Atenolol	6.10	Amitryptiline	8.48
Hydromorphone	6.16	Canrenone	8.48
Sotalol	6.25	Trimipramine	8.55
Codeine	6.30	Brallobarbital	8.56
Dihydrocodeine	6.42	Nelfinavir	8.57
Norcodeine	6.42	Sertraline	8.58
6-Acetylmorphine	6.51	Zuclopenthixol	8.58
Acetaminophen	6.53	Saquinavir	8.61
Olanzapine	6.53	Thioridazine	8.93
Hydrocodone	6.55	Crimidine	8.97
Pseudoephedrine	6.55	Furosemide	8.97
Ephedrine	6.56	Phenolphthalein	9.08
Nadolol	6.60	Alprazolam	9.16
Tubocurarine	6.67	Lorazepam	9.28
Nalbuphine	6.76	Propyphenazone	9.67
Acetazolamide	6.81	Amprenavir	9.80
Ritalinic acid	6.86	Tolbutamide	9.80
Pindolol	6.89	Rhein	10.30
Benzoylecgonine	6.90	Aloeemodine	10.34
Mepivacaine	7.00	Acenocoumarol	10.40
Acebutolol	7.05	Gliclazide	10.49
Timolol	7.05	Warfarin	10.53
Lidocaine	7.06	Glibornuride	10.57
Metoprolol	7.19	Bisacodyl	10.59
Cocaine	7.27	Glibenclamide	10.73
Oxprenolol	7.32	Ritonavir	10.82
Hydrochlorothiazide	7.43	Diazepam	10.87
Venlafaxine	7.45	Lopinavir	10.89
Cocaethylene	7.50	Phenprocoumon	10.91

Bupivacaine	7.76	Diclofenac	11.11
Propranolol	7.76	Coumachlor	11.12
Alprenolol	7.79	Coumatetralyl	11.17
Chlorthalidone	7.80	Efavirenz	11.58
Quetiapine	7.88	Emodine	11.65
Indinavir	7.92	Mefenamic acid	12.05
Torasemide	7.99	Bromadiolone	12.85
Citalopram	8.02	Chlorophacinone	13.78
Levomepromazine	8.02	Cannabidiol	14.15
Nevirapine	8.08	Cannabinol	16.58
Chlordiazepoxide	8.17	delta-9-THC	18.23
Flupenthixol	8.33	delta-8-THC	18.44
Cinchocaine	8.39		

Table 24 – Recorded LC retention times of 87 drugs out of 365 as an example. (The peaks of these recorded LC retention times show a possible overlay as following)

Compared to HPLC, only a rudimentary separation of substances is necessary to detect the analytes at low concentrations. Ionization of mobile phase components (acetonitrile, ammonium formate) and endogenous compounds is the main source of background noise. Contamination of the mass spectra by these compounds potentially hampers the identification expected analytes at low concentrations. The screening and extraction of a sample can be performed for both modes (positive and negative) in less than one hour including the library search which is an acceptable analytical time for a GUS.

Out of all tested compounds (about 365), only 13 (3.7%,amobarbital, acetylsalicylic acid, butalbital, carbromal, coumaphos, ibuprofen, methylphenidate, naproxen, pentobarbital, salicylic acid, secobarbital, spironolactone and thiopental) were not detectable with this LC-MS method. These compounds were identified neither at high therapeutic concentrations nor at low toxic concentrations. Methylphenidate could be detected by its metabolite ritalinic acid, acetylsalicylic acid and salicylic acid by the metabolite gentisic acid. Most of the undetectable drugs were acidic compounds belonging to the class of analgesics or barbiturates. These compounds were undetectable because the ionization efficiency was very low or the normalized collision energy was too high for fragmentation to occur. Generally, these compounds have high therapeutic serum concentrations. As an alternative to this drawback of the method, these substances could, for example, be detected with HPLC-DAD methods described in literature [338].

APCI was preferred to electrospray ionization in order to reduce the risk of ion suppression. This phenomenon affects the formation of the analyte ions during the electrospray process. Sample matrix and co-eluting compounds can contribute to ion suppression. Although ion suppression can have effects on both electrospray ionization and APCI, evidence indicates that the electrospray interface is more impacted [343, 345, 346].

The method of choice in this study to detect and identify compounds was a DDA procedure. As shown in Table 3, compounds can elute at identical retention times. In this case, only the mass spectra of the ion with the highest intensity would be detected. In contrast, ions with low intensity would be lost. To overcome this problem a refractory period was introduced. Based on the average peak width this period was set at 30 seconds. A refractory period longer than 30 seconds can result in a loss of identification of one or more of the compounds with the same molecular mass ion. With a shorter refractory period the method can fail to detect substances eluting with a similar retention time.

Due to different chemical properties of the substances mass spectra (established with pure drugs in aqueous solution) were recorded in the negative as well as in the positive mode. Limit of detection (LOD) values differed for the distinct modes. For example, morphine was better detected in the positive mode, whereas bromadiolone was better detected in the negative mode. For both modes a library was created. At best, substances detected in both polarities could be identified in the two respective libraries. The normalized collision energies of 40.0% in the positive and 35.0% in the negative mode were empirically chosen in order to obtain fragmentation of the compounds. A decrease in the normalized collision energy would yield less fragmentation. Applying higher normalized collision energy would result in lower peak intensities of the fragments because further fragmentation occurs in most cases.

The established library includes for each spectrum the name of the compound, the molecular formula and the molecular ion together with its relative retention time for all 350 substances (the MSMS spectra are in an Xcalibur library which is connected to the library search program). This mass spectral library comprises spectra of about 280 drugs and 85 metabolites (important metabolites for screening in urine) from a large diversity of substance classes. With this procedure, some acidic, neutral as well as basic drugs could be detected and identified. A small application program was developed for the automated identification of unknown compounds with LC-MS.

In order to identify unknown compounds in a serum sample a chromatographic run was performed in each ionization mode. In the next step the developed application program compared each recorded MS-MS spectrum to the reference spectra in the library from the Xcalibur software. With the help of this application program the number of best hits that the unknown spectrum should be compared to could be specified. In the procedure described in this paper, the ten best hits were chosen.

The similarity between the library spectra and the unknown spectra is characterized by the match factor and the reverse match factor. The match factor indicates the correlation between the unknown spectrum and the library spectrum (presence and relative intensities of mass-to-charge ratios). The reverse match factor indicates an inverse search where the presence and the relative intensity of the ions of the library spectrum are compared to those of the unknown spectrum. This parameter ignores the ions present in the unknown spectrum if absent in the reference spectrum.

The match factor and the inverse match factor range between 0 and 1000 with 0 indicating no similarity and 1000 indicating perfect similarity. About 37 compounds newly added to the existing library were tested on other equipments from Thermo Finnigan like a LCQ Deca and a new generation of linear iontrap the Thermo Finnigan LTQ, for comparability of spectra production. All spectra were identical. In our procedure, the threshold was set at 400 for both factors. With this threshold, the best results were obtained. If the maintenance of the equipment follows the recommendations of the manufacturer, this threshold is not varying (experience after routine use of the method since 3 years). For security reasons, this threshold should be approved in each laboratory using other equipments. A higher threshold resulted in a higher LOD together with more specificity of results. In contrast a lower threshold resulted in a lower LOD but also in a higher number of false positives (which is not a big problem when metabolites are present, the whole ion scan including Bruch pieces is consulted, metabolites are present or but a better way in such cases would be to work with MS^3 or MS^4). The following example (Figure 57) shows a run with a serum sample spiked with phenolphthalein, gliclazide, bisacodyl, glibornuride and glibenclamide at a concentration of 1 mg/l. The run was performed in the positive mode.

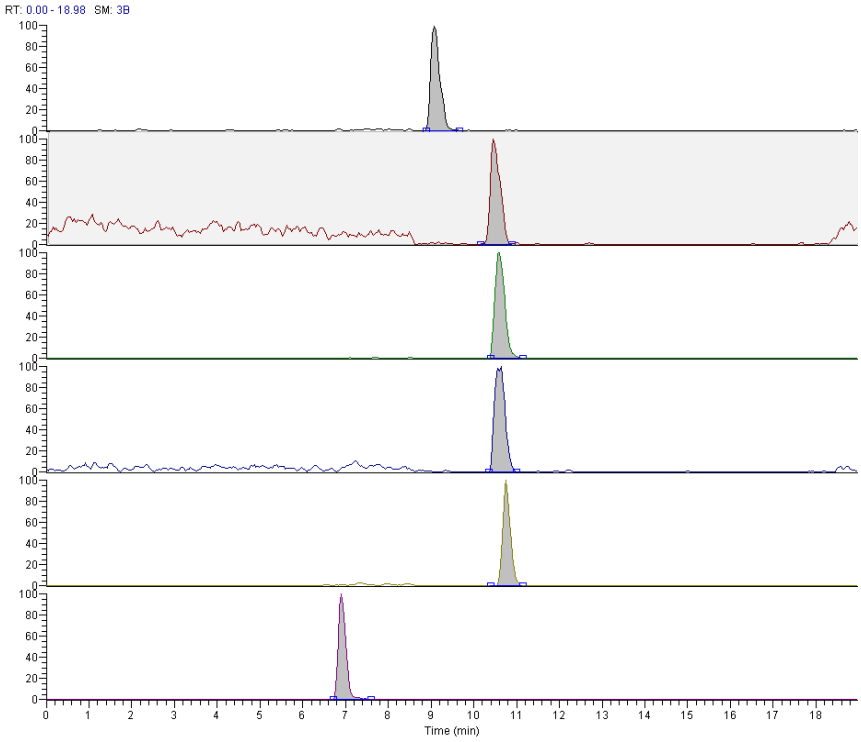


Figure 57 - Chromatogram of 5 substances and the internal standard (D3 benzoylecgonine) all together spiked into a drug free serum (see Material section 2.1) of a 1 mg/l IS 5 mg/l concentration.

Each product ion mass spectrum was subjected to an automated library searching routine against the library spectra. Figure 58 shows the MS-MS spectrum of glibornuride from spiked serum obtained with this procedure compared to the MS-MS spectrum of the library. The match factor of the presented mass spectra was 889, the reverse match factor was 989.

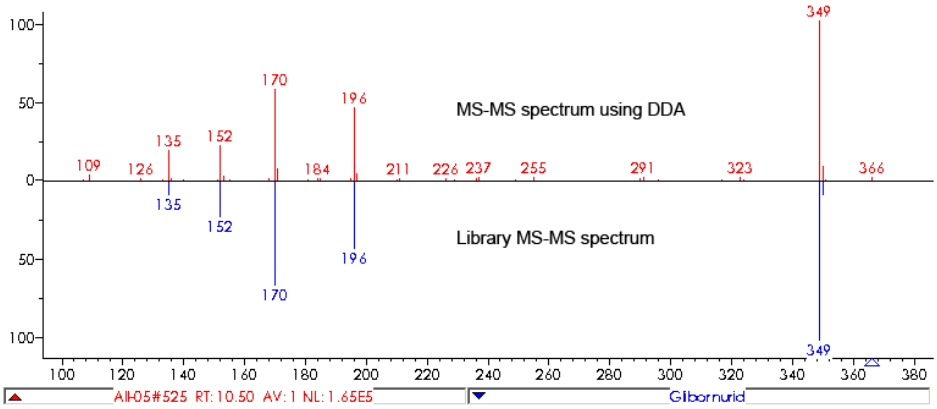


Figure 58 - Comparison of the positive tandem mass spectrometry (MS-MS) spectrum of glibornuride spiked at a concentration of 1mg/l in a serum sample using data-dependent acquisition (DDA) to the MS-MS spectrum of the library.

Compound identification took into account the mass-to-charge ratio of the unknown compound selected before fragmentation. The mass-to-charge ratio had to be within ± 2 m/z of the reference mass-to charge ratio recorded in the library to be considered as a hit. This width of the mass-to-charge ratio window allowed the search for possible isotopes of the compounds (see Figure 60 where several possibilities are proposed for one substance). With a larger window the risk for false positives was increased during routine screening with the described method the search program proposed some false positive for substances arising if many compounds were in the sample. The comparison of the whole ionscan spectre resulted in exclusion of such mistakes. All these pitfalls are documented as addition to a SOP (standard operation procedure for interpretation). A smaller window resulted in false negatives because of differences in mass related to isotopes of the element.

The pseudo-molecular ion (usually protonated in the positive mode, deprotonated in the negative mode) and its fragments were detected and compared to references in the library. Each MS-MS spectrum recorded was derived from one single mass-to-charge ratio (representing the most intense ion of the previous full scan). Other authors used collision-induced dissociation at different voltages to obtain the same information. Mass spectra were acquired by continuously switching between a low and a high orifice voltage throughout the run to obtain both protonated molecular ion (low-voltage scan) and mass spectral fragments (high voltage scan) from the CID in the ion source [7,17]. With the procedure presented in this study it was not necessary to switch between different orifice voltages.

RRT was also considered in the identification procedure. The RRT of the unknown compound had to be within a range of ±5% of the reference RRT (routine experience showed a 1 to 2% variation of the RRT's. This large time-window was chosen because the MS-MS spectra could be obtained within the whole peak width and the refractory period was set at 30 seconds. The RT of the IS was registered in the positive mode with a value of approximately 6.9 min. Variations in RT occur when using different lots of columns with the same adsorption material.

Only if all the parameters were within the fixed areas a hit was reported. In summary the match factor and the reverse match factor had to be above 400, the mass-to-charge ratio had to be ± 2 *m/z* and the RRT had to be within 5% compared to the library parameters (Figure 59). Each MS-MS spectrum, which fulfilled these conditions, was reported. The new program automatically releases a report, which consists of different hits with the substance names together with the parameters mentioned above compared to the ones in the library.

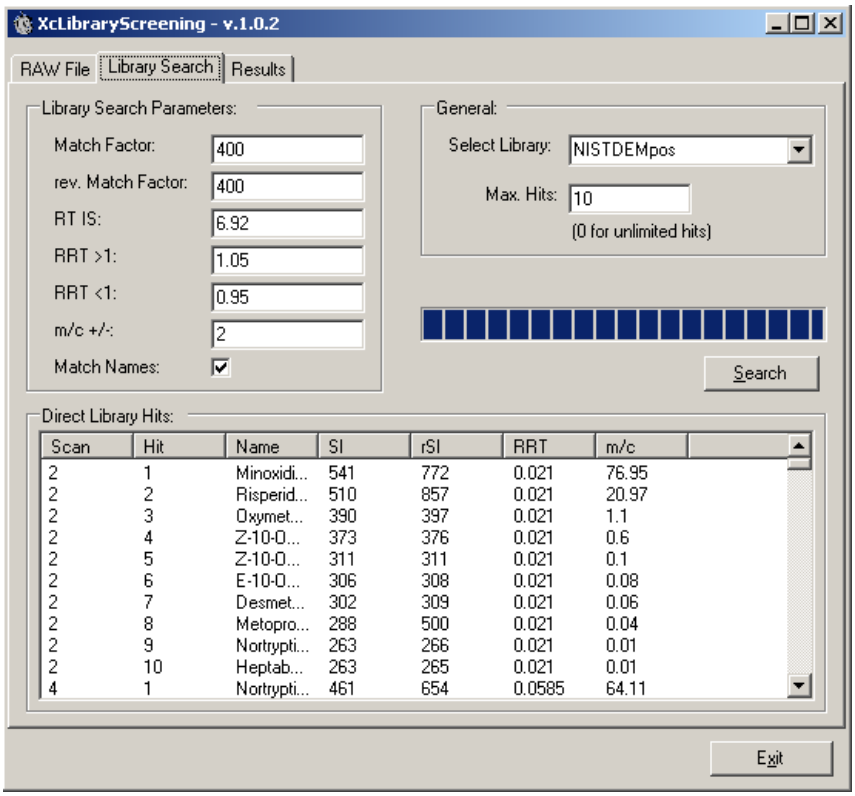


Figure 59 - The library search parameters and the library can be specified. In the described method the threshold of the match factors were set at 400, the relative retention time (RRT) has to be within ± 5% and the maximum deviation of the mass-to-charge was two.

In the above mentioned serum sample all 5 substances (phenolphthalein, gliclazide, bisacodyl, glibornuride and glibenclamide) were identified at a concentration of 1 mg/l with our procedure as an example (concentration in a toxic level for these substances). Figure 60shows the output generated by the application program and the instrument software. The substances were listed in the results file. Gliclazide, bisacodyl, glibornuride and glibenclamide all have similar retention times. However, this was no problem because the refractory period applied enabled to identify the compounds even though they were not chromatographically separated. Importantly, in contrast to single MS CID methods the co-eluting substances did not affect the MS-MS spectra in the presented procedure. Therefore, the analysis of

unknown compounds was more rapid using MS-MS (there was no ion suppression detectable between co eluting substances so far).

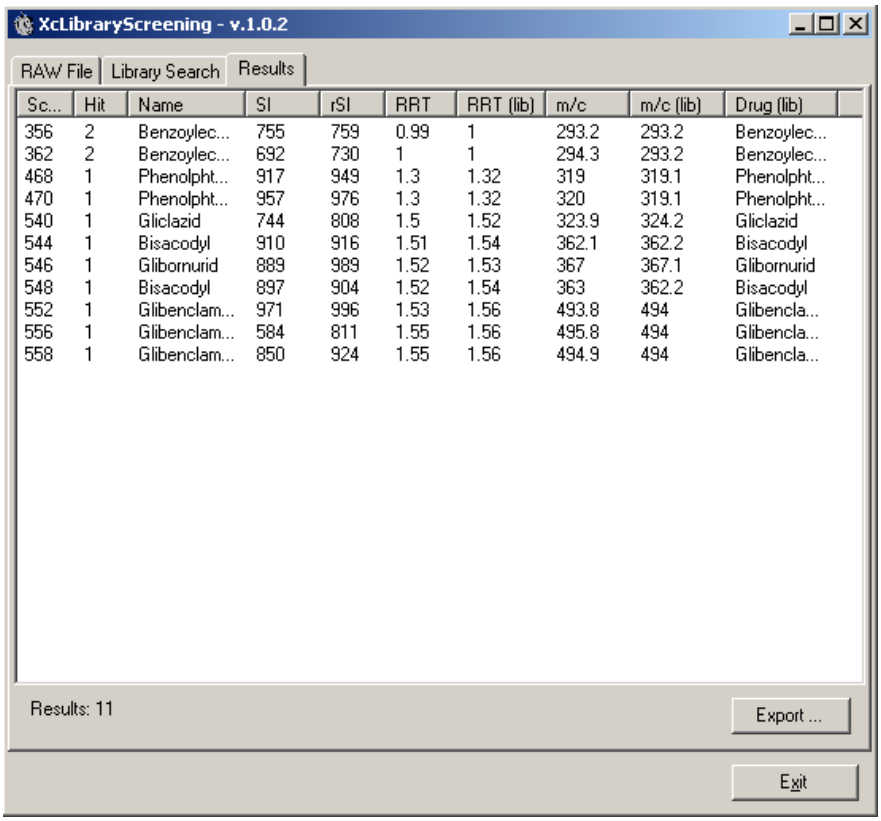


Figure 60 - In this result file, the different hits are presented with their substance name, match factor, reverse match factor, relative retention time (RRT) and mass-to-charge relation compared to the corresponding parameters in the library. The same substance can be found several times with different tandem mass spectrometry (MS-MS) spectra.

Serum samples spiked with decreasing concentrations of the tested drugs were analyzed in order to determine the LOD's. Each concentration of the drugs from 0.005 to 4 mg/l was extracted and analyzed two times. The LOD was defined as the lowest concentration where both runs of each substance fulfilled the mentioned requirements to be identified.

Table 25 shows the different LODs for 87 compounds detected either in the positive and/or in the negative mode. With our procedure all these 87 compounds out of 365 tested for LOD could be detected at high therapeutic drug concentrations or at concentrations in the low toxic range. For drug confirmation measurements in urine all substances like opiates, amphetamines, cocaine metabolites have LOD's beyond the SAMSHA cutoff for chromatographic confirmations. THC-carbonic acid PCP and LSD cannot be confirmed with this method. The LOD was ≤ 100 µg/l for 67% of the compounds. Most of the drugs were better detected in the positive mode, especially compounds with chemical structures of amines such

as neuroleptics, opioids and anti-depressants. In the negative mode a lower LOD was seen with molecules containing acidic sites like diclofenac.

	LOD (µg/l)			LOD (µg/l)	
	positive	negative		positive	negative
6-Acetylmorphine	100		Hydrochlorothiazide		4000
Acebutolol	100		Hydrocodone	50	
Acenocoumarol		25	Hydromorphone	50	
Acetaminophen	500		Imipramine	500	
Acetazolamide		2000	Indinavir	25	250
Aloeemodine		50	Levomepromazine	500	
Alprazolam	50		Lidocaine	250	
Alprenolol	250		Lopinavir	50	25
Amiloride	100		Lorazepam	100	100
Amitryptiline	25		Mefenamic acid	250	250
Amprenavir	50	50	Mepivacaine	100	
Atenolol	25		Metoprolol	250	
Benzoylecgonine	50		Morphine	50	
Bisacodyl	100		Nadolol	50	
Brallobarbital		1000	Nalbuphine	50	
Bromadiolone	1000	250	Nelfinavir	25	100
Bupivacaine	250		Nevirapine	50	
Cannabidiol	100		Norcodeine	100	
Cannabinol	250		Olanzapine	250	
Canrenone	50		Oxprenolol	250	
Chlordiazepoxide	100		Phenobarbital		2000
Chlorophacinone		50	Phenolphthalein	50	
Chlorthalidone	1000		Phenprocoumon	500	50
Cinchocaine	25		Pindolol	100	
Citalopram	100		Propranolol	100	
Cocaethylene	500		Propyphenazone	50	
Cocaine	100		Pseudoephedrine	100	
Codeine	100		Quetiapine	50	
Coumachlor	500	50	Rhein		50

Coumatetralyl	1000	100	Ritalinic acid	500	
Crimidine	250		Ritonavir	50	
Delta-8-THC	100		Saquinavir	10	
Delta-9-THC	100		Sertraline	250	25
Diazepam	250		Sotalol	25	
Diclofenac	1000	500	Thioridazine	500	
Dihydrocodeine	25		Timolol	100	
Efavirenz		100	Tolbutamide		250
Emodine		50	Torasemide	100	
Ephedrine	500		Trimipramine	250	
Flupenthixol	100		Tubocurarine	100	
Furosemide		100	Venlafaxine	250	
Glibenclamide	25	50	Warfarin	250	25
Glibornuride	100	100	Zuclopenthixol	100	

Table 25 – LODs in serum determined for 87 drugs out of 365 (as an example) applying negative and positive ionization mode.

The following examples illustrate the application of the DDA LC-MS-MS system to clinical investigations (Table 26). The results of serum and urine sample analysis using SPE-LC-MS-MS were compared to the results obtained with a conventional STA strategy (including a combination of immunoassays, HPLC (Remedi) and LC-MS). REMEDi (Biorad) is an HPLC-based broad-spectrum drug profiling system. It is used to detect and identify basic and neutral drugs and their metabolites in serum and urine samples of patients) [342]. To identify benzodiazepines an additional run has to be performed on a special Remedi System. This conventional STA strategy has been applied for several years in the laboratory of the University Hospital Basel.

Projects

No.	Conventional STA	DDA LC-MS-MS
1 (Serum)	Benzodiazepines (CEDIA)	Oxazepam, Oxazepam Metabolite, Desmethyldiazepam
	Metoclopramide (REMEDi)	Metoclopramide
2 (Serum)	Benzodiazepines (CEDIA)	Oxazepam, Oxazepam Glucuronide
	Ranitidine (REMEDi)	Ranitidine
	Lidocaine, Lidocaine Metabolite (REMEDi)	Lidocaine
	Atracurium (REMEDi)	nl
	Dipyrone (REMEDi)	nl
	nd	Metoclopramide
3 (Serum)	Antidepressants (DRI), Trimipramine Meta-bolite (REMEDi)	Trimipramine, Desmethyltrimipramine
	Methadone (CEDIA, REMEDi)	
	Benzodiazepines (CEDIA)	Methadone
		Diazepam, Desmethyldiazepam
4 (Serum)	Citalopram (REMEDi)	nd
	Clozapine, Clozapine Metabolite (REMEDi)	Clozapine, N-Desmethylclozapine, Clozapine-N-Oxide
	nd	Amisulpiride
	nd	Atenolol
5 (Serum)	THC-Carbonic acid (CEDIA)	nd
	Fluoxetine, Fluoxetine Metabolite (REMEDi)	Fluoxetine
6 (Serum)	Benzoylecgonine (EMIT, REMEDi), Cocaine (REMEDi)	Benzoylecgonine
7 (Serum)	Opiates (CEDIA), Pethidine, Pethidine Meta-bolite (REMEDi)	Pethidine
	Benzodiazepine (CEDIA)	
	Dipyrone (REMEDi)	Diazepam
		nl
8 (Serum)	Antidepressants (DRI), Amitryptiline (REMEDi)	Amitryptiline, 10-OH-Amitryptiline, Nortryptiline
		Zolpidem
	Zolpidem (REMEDi)	nd
	Salicylates (EMIT)	
9 (Serum)	Quetiapine, Quetiapine Metabolite (REMEDi)	Quetiapine
	nd	Lamotrigine
10 (Serum)	Benzodiazepines (CEDIA)	Lorazepam

11 (Serum)	Opiates (CEDIA) Benzoylecgonine (CEDIA, REMEDi), Cocaine (REMEDi)	Codeine-6-glucuronide Benzoylecgonine
12 (Serum)	nd	Mefenamic acid
13 (Urine)	Opiates (CEDIA), Morphine, Morphine Meta-bolites, Codeine Metabolites (all REMEDi and LC-MS), Codeine (LC-MS)	Normorphine, Morphine, Codeine-6-Glucuronide, Codeine, Norcodeine
14 (Urine)	Opiates (CEDIA), Morphine, Codeine Metabolite, Codeine (all REMEDi) Atenolol (REMEDi)	Normorphine, Morphine-Glucuronide, Codeine-6-glucuronide, Codeine Atenolol
15 (Urine)	Benzoylecgonine (CEDIA, REMEDi), Cocaine (REMEDi) Mepivacaine	Benzoylecgonine, Cocaine, Cocaethylene Mepivacaine
16 (Urine)	Citalopram Levomepromazine, Levomepromazine Meta-bolite Clozapine, Clozapine Metabolite Tiapride	Citalopram Levomepromazine Metabolites Clozapine, Clozapine Metabolites Tiapride
17 (Urine)	Methadone (CEDIA), EDDP (CEDIA, REMEDi) Opiates (CEDIA), Dihydromorphine, Hydro-morphone (both REMEDi), Hydrocodone (REMEDi, LC-MS), Dihydrocodeine, Morphine Metabolite (LC-MS) Benzodiazepines (CEDIA), Desmethyl-diazepam, Temazepam (REMEDi) THC-Carbonic acid (CEDIA)	EDDP Hydrocodone, Dihydrocodeine, Dihydrocodeine-6-Glucuronide, Norcodeine Desmethyldiazepam, Temazepam-glucuronide, Oxazepam, Oxazepam-glucurondie 11-nor-delta-THC-COOH
18 (Urine)	Mirtazapine Metabolite (REMEDi)	Mirtazapine, Desmethylmirtazapine
19 (Urine)	Venlafaxine, Venlafaxine Metabolite (REMEDi) Benzodiazepines (CEDIA) Oxazepam (REMEDi) nd nd nd	Venlafaxine, Desmethylvenlafaxine Oxazepam, Oxazepamglucuronide Mirtazapine Zolpidem Metoprolol
20 (Urine)	Antidepressants (CEDIA), Amitryptiline, Nortryptiline (REMEDi) Benzodiazepines (CEDIA)	Nortryptiline, Hydroxynortryptiline Bromazepam, Hydroxybromazepam

22 (Urine)	Amphetamines (CEDIA), MDMA (REMEDi, LC-MS), Ephedrine, MDA, Amphetamine, Metamphetamine (all LC-MS)	Ephedrine, MDMA
	Benzoylecgonine (CEDIA, REMEDi)	
	Benzodiazepines (CEDIA)	Benzoylecgonine
	Methadone, EDDP, (CEDIA)	
	nd	EDDP
		Lidocaine
23 (Urine)	Lidocaine, Lidocaine Metabolite (REMEDi)	Lidocaine
	Mirtazapine Metabolite (REMEDi)	Mirtazapine
	Chlorprothixene Metabolite	Chlorprothixene
	Opiates (CEDIA), Morphine (REMEDi)	Morphine, Codeine, Codeine Metabolite
	nd	Paroxetine
	nl	Furosemide
24 (Urine)	Amphetamines (CEDIA), MDMA, MDA (REMEDi)	MDMA
	Cocaine, Benzoylecgonine (REMEDi)	
	Benzodiazepines (CEDIA), Flurazepam Metabolites	Benzoylecgonine
		Flurazepam, Hydroxyethylflurazepam, Hydroxybromazepam
25 (Urine)	Benzoylecgonine (CEDIA, LC-MS), Cocaine (LC-MS)	Benzoylecgonine
	Benzodiazepine (CEDIA), Temazepam, Desmethyldiazepam (REMEDi)	
	Opiates (CEDIA)	Oxazepam, Temazepam, Temazepamglucuronide
	THC-Carbonic acid (CEDIA)	nd
		nd
26 (Urine)	Olanzapine (REMEDi)	Olanzapine
	Quetiapine Metabolite (REMEDi)	nd
	Amphetamines (CEDIA) MDA, MDMA (REMEDi, LC-MS), MDEA (LC-MS)	MDA, MDMA, MDEA
	Benzodiazepines (CEDIA), Flurazepam Metabolite (REMEDi)	
	Carbamazepine (REMEDi)	nd
		Carbamazepine

Table 26 – Comparison of DDA LC-MS-MS procedure to conventional STA Technique (Immunoassay, LC-MS and REMEDi) of serum and urine samples

In addition, the different opiates, amphetamines, benzoylecgonine, cocaine, ethylcocaine, buprenorphine, norbuprenorphine, and additional opioides were analyzed on a validated LCMS system with methods

adapted for each drug class (lower LOD's between 5 to 20ng/ml. These methods are not only used for confirmation but differenciation and quantification of the substances). Twelve serum and fourteen urine samples of expected intoxicated individuals and drug addicts were analyzed (Table 26). With the new method, urine samples were treated like serum samples (automated extraction followed by chromatography).

In addition to the drugs identified by the conventional STA, the newly developed DDA LC-MS-MS system also found metoclopramide (case 2), trimipramine (case 3), amisulpride, atenolol (case 4), amitryptiline metabolite, nortryptiline (case 8), lamotrigine (case 9) and mefenamic acid (case 12) in serum samples (confirmed by HPLC). In contrast, the lidocaine metabolite (MEGX), atracurium (both case 2), dipyrone (metamizole, case 2 and 7), fluoxetine metabolites (case 5), pethidine metabolites (case 7) and quetiapine metabolites (case 9) were not detected most probably because the MS data of these substances were not included into the library.. The LC-MS-MS procedure failed to detect citalopram (case 4), THC (THC or THC-carbonic acid, case 5), cocaine (case 6 and 11) and salicylates (case 8). THC and cocaine were at a concentration beyond their LODs. Salicylates are not detectable with this system as described above.

In urine samples, a number of drugs, which were not found by the conventional STA procedure, were detected by the DDA approach including oxazepam (case 17), zolpidem, metoprolol (both case 19), lidocaine (case 22), codeine, paroxetine, furosemide (case 23), and bromazepam metabolite (case 24). On the other hand amitryptiline (case 20), amphetamine, metamphetamine (case 22), THC-carbonic acid (case 25), desmethyldiazepam (case 25), quetiapine metabolite (case 26) and flurazepam metabolite (case 26) were missed by the newly developed system. 7-hydroxy-quetiapine was not detected because its MS data was not included in the library.

In general, the new system identified – with some exceptions - the same drugs as the conventional STA and some additional substances. With the developed LC-MS-MS procedure, basic, neutral as well as acidic substances can be identified within the same system. However, some of the acidic compounds should be analyzed with a different procedure.

Importantly, with the new method, drugs were identified in two runs and hydrolysis of glucuronides was not necessary (except for the screening of benzodiazepines). Furthermore, time-consuming dilutions of samples can be avoided. The presented approach was robust and the information content was high. Substances from different groups (amphetamine-derived designer drugs, antidepressants, benzodiazepines, cocaine, opiates, antidiabetics, neuroleptics etc.) were detected and identified.

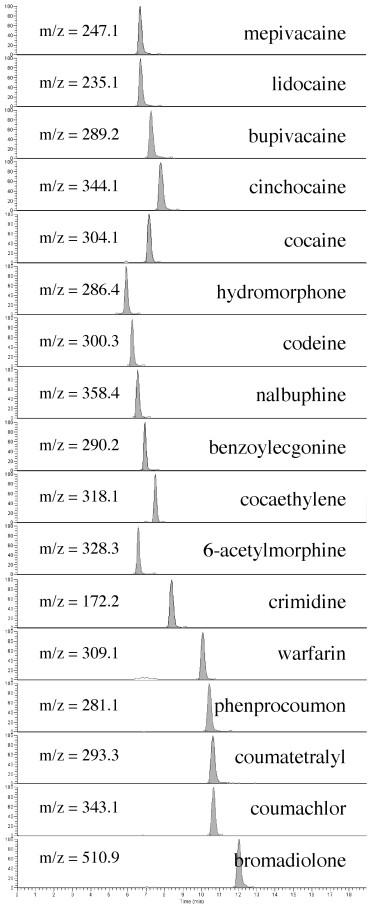


Figure 61 – Extracted ion chromatograms (XICs) of aqueous solutions spiked with compounds. 17 compounds are displayed in the positive mode.

The results of the analysis of 25 urine samples using the new method were compared to those obtained using an established GC-MS screening method in an external laboratory. Table 27 shows the results of these comparisons. Unfortunately, the sample volumes were lower than described for this GC-MS STA, so that a few drugs with low concentrations could not be detected. For the same reason, the screening for acidic compounds could not be performed.

No. Urine	Conventional GUS, GC-MS (Maurer, Homburg)	DDA LC-MS-MS
1	Morphine	Morphine, Normorphine
	Papaverine	nd
	Codeine	Codeine, Codeineglucuronide
	Amphetamine	nd (masked by very high concentration of a different substance)
	Ephedrine/Norephedrine	Ephedrine, Pseudoephedrine
2	-----	-----

3	Diazepam	Oxazepam, Temazepam + Glucuronide, Desmethyldiazepam
	Flurazepam	nd
	Morphine	Morphine, (Morphineglucuronides
	Apomorphine	nd
	Midazolam + metabolites	Midazolam, α-Hydroxymidazolam
		7-Aminoflunitrazepam
		Mefenamic Acid
4	---	Lormetazepam
5	Citalopram	Citalopram, Desmethylcitalopram
		Naltrexone
		Temazepamglucuronide
6	Diazepam + metabolites	Oxazepam, Temazepam, Temazepamglucuronide
	Doxepin + metabolites	Doxepin, Desmethyldoxepin
	Lorazepam	nd deglu
	Triclosan	nl
		Dibenzepine
7	Venlafaxine + metabolites	Venlafaxine, Desmethylvenlafaxine
	Mirtazapine + metabolites	Desmethylmirtazapine
	Lorazepam	nd deglu
8	Triclosan	nl
9	Methadone	Methadone
	Diazepam + metabolites	Oxazepamglucuronide, Temazepamglucuronide, Desmethyldiazepam
	Mirtazapine	Mirtazapine
10	---	---
11	Paracetamol	nd deglu
	Tramadol	Tramadol (traces)
	Morphine	Morphine
	Sertraline	Masked by high concentrations of EDDP
	Methadone	Methadone, EDDP
	Diazepam + metabolites	Oxazepam, Oxazepamgluc., Temazepamgluc., Demethyldiazepam
	Fluconazole	nl
		Metoclopramide

12	Paracetamol	nd deglu
	Metoclopramide	Metoclopramide
		Benzoylecgonine
		Oxazepam
13	Lorazepam	nd deglu
	Ticlodipine	nl
	Lidocaine	nd
	Paracetamol	nd deglu
	Methadone + metabolites	Methadone, EDDP
	Diazepam + metabolites	Oxazepam, Temazepamglucuronide
	Metoprolol	Metroprolol
14	Morphine	Morphine, Morphineglucuronides
	Amphetamine	nd
	Diazepam + metabolites	Oxazepam, Oxazepamglucuronide, Temazepam, Temazepam-glucuronide .
	Methaqualone	Methaqualone
		Carbamazepine
		Methylphenidate
15	Paracetamol	nd deglu
	Diazepam + metabolites	Oxazepamglucuronide., Temazepam and Temazepamglucuronide
	MDMA + metabolites	MDMA, MDA
	Codeine	Codeine, Codeineglucuronide
	Morphine + metabolites	Morphine, Normorphine
		Noscapine
		Chlordiazepoxide
		7-Aminoflunitrazepam
16	Morphine + metabolites	Morphine, (6-Acetylmorphine, Diacetylmorphine, Normorphine)
	Diazepam + metabolites	Diazepam, Desmethyldiazepam, Oxazepam
	Trimipramine + metabolites	Desmethyltrimipramine, Trimipramine, Hydroxy-Trimipramine
	Midazolam + metabolites	4-Hydroxy- Midazolam
		7-Aminoflunitrazepam
		Apomorphine
17	Oxazepam	nd

18	Oxazepam	Oxazepam, Oxazepamglucuronide
	Lidocaine	nd (to low signals of the internal standard, no rerun possible)
	Lorazepam	nd deglu
	Methadone + metabolites	Methadone, EDDP
	Trometamol	nl
19	Paracetamol	nd deglu
	Morphine	nd
	Methadone + metabolites	Methadone, EDDP
	Codeine	nd
		Venlafaxine
20	Paracetamol	nd deglu
	Mirtazapine + metabolites	Mirtazapine, Desmethylmirtazapine
	Flurazepam	N-1-OH-Ethylflurazepam
	Clotiapine	Clotiapine
	Betablocker	Atenolol
21	Morphine	
22	Morphine	Morphine, Morphine-6-Glucuronide, Normorphine
	Methadone + metabolites	Methadone, EDDP
	Diazepam + metabolites	Oxazepam + Gluc., Temazepamgluc., Diazepam, Desmethyldiazepam, Diazepam
	Tramadol	nd
		Venlafaxine, Desmethylvenlafaxine
		Olanzapine
23	Clozapine + metabolites	Clozapine, Desmethylclozapine
	Lorazepam	Lorazepam
24	Lidocaine	nd
	Midazolam + metabolites	α-Hydroxy-Midazolam, 4-OH-Midazolam, Midazolam
	Fentanyl	nd (in general not detected at low concentrations)
	Atracurium	nl
25	Amitriptyline	nd (confirmation negative: HPLC, IA)
	Methadone + metabolites	Methadone, EDDP
	Midazolam + metabolites	α-Hydroxy-Midazolam
		Benzoylecgonine (confirmed)

Table 27 – Comparison of DDA LC-MS-MS procedure to conventional GUS Technique (GC-MS) of urine samples. Nicotine and caffeine are not listed.

In the 25 urine samples, 83 corresponding or different substances were found by the GC-MS screening method (urines tested negative for any substance were defined as 1 “substance” result). 50 results (60%) were in agreement with the new LC-MS-MS method. In 6 cases, drugs found by GC-MS could not be confirmed because they were not included in the LC-MS-MS library (7.2%). Ten further findings (12.1%) obtained by GC-MS were not in agreement with the presented procedure because the glucuronides were not hydrolysed for the measurement with the LC-MS-MS method at that time. Therefore two substances (paracetamol and lorazepam) were missed in all these samples. The glucuronides of these two substances were not included in the library.

Three substances (3.6%), positive by GC-MS and negative by LC-MS-MS, could not be confirmed as positive by additional analysis methods (immunoassays, HPLC). Three substances detected by LC-MS-MS were negative by GC-MS (3.6%), maybe due to the too low concentrations. Two substances could not be found by LC-MS-MS because they were masked by other substances in high concentrations having the same retention time (2.4%). For example, sertraline in low concentrations was masked by EDDP in very high concentrations. Nine additional substances all at low concentrations (not in the toxic range) were not found by LC-MS-MS (10.8%).

With the new LC-MS-MS method, 120 substances (xx drugs and xx metabolites) were found in the 25 urine samples. Mefenamic acid and 7-aminoflunitrazepam in sample 3, lormetazepam in sample 4, naltrexone and temazepam in sample 5, dibenzepine in sample 6, metoclopramide in sample 11, benzoylecgonine and oxazepam in sample 12, carbamazepine and methylphenidate in sample 14, noscapine, chlordiazepoxide and 7-aminoflunitrazepam in sample 15, 7-aminoflunitrazepam and apomorphine in sample 16, venlafaxine in sample 19 and sample 22, olanzapine and venlafaxine in sample 22, benzoylecgonine in sample 25 were only detected by the new LC-MS-MS method, because – as already mentioned - the sample volume for the GC-MS screening was too low.

Conclusions

In this study, it was demonstrated that the DDA LC-MS-MS screening method seems suitable for routine measurements of serum and urine samples. The described procedure is fully automated (from the extraction to the detection of a drug) and easy to handle. The method was highly specific because compounds were detected and identified by their retention times, the mass to charge ratio of their molecular ions and fragments. Rapid identification in screening experiments was achieved by the creation of the small application program. With the method presented here, the analysis and interpretation of serum or urine samples could be performed in less than one hour. The constructed library comprises more than 400 spectra with the corresponding relative retention times of more than 350 compounds. Most of the compounds were detected at therapeutic concentrations. The matrix effects appeared to be negligible.

However, some pitfalls appeared when substances with the same molecular weight eluted at the same retention time. In such a case, only the substance dominating the mass spectrum might be identified and the other substance with the less intensive peak might be missed. For example, clozapine in a range over

2 mg/l could mask citalopram at a concentration of about 0.1 mg/l. In conclusion, for routine screening, the combination of SPE, LC and APCI-MS seems to represent an attractive alternative for the analysis of samples from suspected intoxicated individuals or drug addicts, if combined with HPLC-DAD.

6. Conclusions and Outlook

This thesis deals with the prediction and modulation of transport, pharmacokinetics, and the effects of drugs and implications in drug discovery, drug safety, and clinical practice. A large part of the work concerns the development and application of machine learning tools for the prediction and analysis of drug properties.

The models presented for P-gp interaction are robust and accurate enough to find application in the real world. The data for which these models were built are a survey of the information available in literature. Not all compounds are drug-like compounds and accuracy for pharmacological purposes may be improved by restricting the set to such substances. More insight into the complex and as of yet poorly understood biology of P-gp may be gained by employing other analytical methods such as the ones described in this thesis.

Data for the cytochrome P_{450} models was based on a library of FDA approved substances. This prior focus certainly contributed much to the high accuracy of the models. The CYP system is highly connected, specificities overlap greatly, and the many possible ways of interaction make it difficult to find structural requirements that are specific to any CYP isoform or mode of interaction. Also, activities are not only derived from *in vivo* data in humans but also from animal experiments and *in vitro* assay. While the CCRs of the models are proof to the validity of this approach, a purified data set using only one class of analysis may yield even better results.

Analysis of structure-activity relationships for general classes of ADRs revealed distinct structural requirements for all four classes examined. It was shown how permeability and solubility affects the occurrence of ADRs in the CNS, liver, and kidney, and how structural complexity influences the allergic potential. While emphasis of this study is clearly on ADRs, the results may contribute to an understanding of how compounds reach their target organs by using ADRs as surrogate marker. Comparison with the data on P-gp interaction from the first study in this thesis indicated that P-gp efflux may not protect from CNS ADRs as much as could be expected. Refinements could certainly be made by including metabolites and by applying any of the other ML methods discussed above.

In the study of P-gp and BCRP SNPs, we tried to predict the risk for developing IBD for subjects with P-gp or BCRP. In this study in healthy controls and IBD patients from the Swiss population, trends were seen for SNPs and haplotypes involving BCRP C421A. This in itself is of great interest for the Swiss medical community. However, the study suffers from a small sample size. Increasing the number of patients would most likely show statistically significant results for some of the genetic polymorphisms discussed. Furthermore, considering the physiological role of these efflux pumps, an understanding of the prevalence of polymorphisms with decreased activity can increase safety of drug therapy not only in IBD patients but also the general Swiss population. Still, it is evident that for multifactorial diseases such as IBD no single

genetic marker for susceptibility or disease phenotype will likely be identified, but the three SNPs and the haplotypes derived from them may prove useful in the prediction of such parameters in the future.

In the study of St. John's Wort in a single patient suffering from Crigler-Najjar syndrome type II, clinically as well as statistically significant reduction of plasma bilirubin was seen, leading to an evident decrease in jaundice and increase in quality of life for the patient. This supports the prediction that the constitutionally decreased activity of UGT 1A1 can be enhanced by translational activation (in this case, via pregnane X receptor activation). If other patients with CN II (or even just phenobarbital sensitive subtypes) benefit from this finding, remains to be seen. Given the rarity of the disease, it will be difficult to conduct larger scale studies.

The evaluation of the prototype device for the pulsatile transdermal delivery of nicotine was successful as judged by the plasma levels. As of now, the device itself is cumbersome and requires an external computer to control it. Efforts are being made by the developing company to improve on size and handling. How this form of delivery compares to conventional means of nicotine substitution and what other compounds the device may deliver in a similarly safe manner remains to be seen in further studies.

Lastly, the automated general unknown screening system developed in the final (isolated) project has successfully been in use for several years at the University Hospital of Basel, handling the majority of GUS tasks in blood and urine. Minor revisions were made to port it to the next generation LC/MS machines and recently it has been extended to not only report qualitative results (i.e. the presence of toxic compounds) but also quantitative results (estimating the absolute amounts of substances).

7. Appendix

7.1. Source code

7.1.1. Retrieval of structure information from PubChem in Ruby

The following script program performs a query over a hypertext transfer protocol (http) connection to the PubChem database (http://pubchem.ncbi.nlm.nih.gov/) of the National Center for Biotechnology Information (Bethesda, MD, USA). It takes the Compound Identification Number (CID), which is unique for each compound in the database, and retrieves an extensible markup language (XML) file containing systematic names, fingerprints, and several descriptors (subject to change without notice) along with the structure represented as canonical and / or isomeric SMILES codes. The script then parses the information and returns it in tabular format suitable for import in any spreadsheet program or database system. It was written in Ruby (Version 1.8.6, http://www.ruby-lang.org/) and uses the XML conformant parser REXML (Version 3.1.7.3, http://www.germane-software.com/software/rexml, also part of the Ruby standard library).

```
#!/usr/bin/ruby -w

# fetch.rb
#
# retrieves compound summary from PubChem
#
# Date: 28-APR-2009
# Ruby 1.8.6
#
# syntax:
# ruby fetch.rb CID [-toCSV] [-headers]
#
# CID: PubChem Compound ID
# -toCSV: dump result on single line with field separated by tabs (optional)
# -headers: dump headers for CSV file with fields separated by tabs (optional)
#
#

require 'net/http'
require 'rexml/document'

if ARGV.length == 0 then
  puts "need at least one argument (CID)"
  exit
end

asCSV = false
dumpHeaders = false
ARGV.each do |arg|
  arg = arg.upcase
  if arg == "-TOCSV" then
    asCSV = true
  end
  if arg == "-HEADERS" then
    dumpHeaders = true
  end
end
```

```
url = 'http://pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?cid=' + ARGV[0] +
'&disopt=SaveXML'
xml_data = Net::HTTP.get_response(URI.parse(url)).body
doc = REXML::Document.new(xml_data)
titles = []
links = []
doc.elements.each('PC-Compound/PC-Compound_props/PC-InfoData/PC-InfoData_urn/PC-Urn/PC-
Urn_label') do |e|
  titles << e.text
end

doc.elements.each('PC-Compound/PC-Compound_props/PC-InfoData/PC-InfoData_value/*') do
|e|
  links << e.text
end

if asCSV then
  puts titles.join("\t") if dumpHeaders
  puts links.join("\t")
  else
    titles.each_with_index do |title, idx|
      print "#{title} => #{links[idx]}\n"
    end
end
```

7.1.2. Calculation of chemical similarity using Tanimoto's coefficient in Java

This program written in Java (Java 2 Platform, Standard Edition, Version 1.4.2, http://www.sun.com/java, Sun Microsystems, Santa Clara, CA, USA) calculates the Tanimoto coefficient of similarity (see 4.1.6.2) based on a vector of numerical features for every instance. Data is supplied in the form of a comma separated file (CSV) using the tabulator character as a field delimiter. It returns the matrix of similarity coefficients for all instances and calculates the overall similarity as described previously.

```
/**
 * SimilarityCalculator
 * calculates Tanimoto coefficient
 *
 * Date: 10-DEC-2008
 * Java 1.4
 */

import java.io.BufferedReader;
import java.io.FileNotFoundException;
import java.io.FileReader;
import java.io.IOException;
import java.util.StringTokenizer;
import java.util.Vector;


public class SimilarityCalc {
    double[][] data;
    char separator = '\t';

    public SimilarityCalc( String fileName, boolean skipHeaders, boolean dumpMatrix )
throws FileNotFoundException, IOException {
        System.out.print( "reading " + fileName + " ... " );
        BufferedReader reader = new BufferedReader( new FileReader(fileName) );
        String line = reader.readLine();
        if( line!=null && skipHeaders ) // suppress headers row?
            line = reader.readLine();

        // count cols
        int colCount = 1;
        for( int i=0; i<line.length(); i++ )
```

```
            if( line.charAt(i) == separator )
                colCount++;

        // because number of lines is unknown at this point,
        // store every row of the matrix in a vector and convert later
        Vector v = new Vector();
        while( line!=null ) {
            double[] row = new double[colCount];
            StringTokenizer tokenizer = new StringTokenizer( line, ""+separator );
            for( int i=0; i<colCount; i++ )
                row[i] = Double.parseDouble( tokenizer.nextToken() );
            v.addElement(row);
            line = reader.readLine();
        }
        data = new double[v.size()][colCount];
        for( int i=0; i<v.size(); i++ )
            data[i] = (double[]) v.elementAt(i);

        System.out.println( v.size() + " line(s). done." );

        double[][] matrix = tanimotoMatrix();
        if( dumpMatrix )
            dumpArray( matrix );
    }

    // dump array to console as CSV readable data
    private void dumpArray( double[][] array ) {
        for( int i=0; i<array.length; i++ ){
            for( int j=0; j<array[i].length; j++ ) {
                if( i==j ) // identity.
                    System.out.print( "1.0\t" );
                else
                    System.out.print( array[i][j] + "\t" );
            }
            System.out.println(  );
        }
    }

    protected double normalizeArray( double[][] array, int count ) {
        double result = 0.0;
        for( int i=0; i<array.length; i++ )
            for( int j=0; j<array[i].length; j++ )
                result+=array[i][j];

        return result / count;
    }

    protected double[][] tanimotoMatrix() {
        double[][] result = new double[data.length][data.length];
        int count=0;
        for( int i=0; i<data.length; i++ ) {
            for( int j=i+1; j<data.length; j++ ) {
                count++;
                double index = 0;
                for( int k=0; k<data[i].length; k++ ) {
                    index += tanimotoDistance( data[i][k], data[j][k] );
                }
                result[i][j] = index / data[j].length;
            }
        }

        System.out.println( "Index: " + normalizeArray(result, count) );

        return result;
    }

    protected double tanimotoDistance( double a, double b ) {
        if( a==0.0 && b==0.0 ) return 0.0; // avoid NaN
        return (a*b)/(a*a+b*b-a*b);
```

```
    }

    public static void main(String[] args) {
        if( args.length < 1 ) {
            System.out.println( "must supply file name" );
            System.exit(-1);
        }

        boolean skipHeaders = false;
        boolean dump = false;
        for( int i=1; i<args.length; i++ ){
            String command = args[i].toUpperCase();
            if( command.equals("-SKIPHEADERS") )
                skipHeaders = true;
            if( command.equals("-DUMP") )
                dump = true;
        }

        try {
            new SimilarityCalc( args[0], skipHeaders, dump  );
        } catch( Exception e ) {
            e.printStackTrace();
        }
    }
}
```

7.1.3. Support Vector Machine grid search in Java

This Java program implements a grid search for support vector machine learning parameters cost (C) and γ (gamma) as previously described (see 4.2.7.3). It uses the Weka toolkit (Version 3.6, http://www.cs.waikato.ac.nz/~ml/weka/, University of Waikato, Hamilton, New Zealand) for cross-validation (see 4.2.2.5) and feature selection (see 4.2.8) [177] and LibSVM as a faster and more versatile implementation of SVM learning algorithms (Version 2.89, http://www.csie.ntu.edu.tw/~cjlin/libsvm) as described by the authors [347]. Integration of LibSVM with the Weka toolkit was achieved with WLSVM, written by Yasser EL-Manzalawy and Vasant Honavar (http://www.cs.iastate.edu/~yasser/wlsvm, Iowa State University, Ames, IO, USA).

```
/*
 * SVM grid search
 * Weka 3.4.13
 * libsvm as wlswm
 *
 * last rev: 17-APR-09
 */

package svmgridsearch;

import java.io.BufferedReader;
import java.io.FileReader;
import java.io.IOException;

import java.util.Random;
import java.util.Vector;
import weka.attributeSelection.BestFirst;
import weka.attributeSelection.CfsSubsetEval;
import weka.classifiers.Evaluation;
import weka.classifiers.functions.SMO;
import weka.classifiers.meta.AttributeSelectedClassifier;
import weka.core.Instances;
import wlsvm.WLSVM;
```

```
/**
 *
 * @author lix
 */
public class SVMGrid {
    private double[][] values;
    private double[][] labels;
    private double[][] results;
    private int svm = 0;
    private int kernel = 2;
    private int c_count = 10;
    private double c_start = -5;
    private double c_step = 2.0;
    private double gamma_start = -15;
    private int gamma_count = 10;
    private double gamma_step = 2.0;
    private double variance_threshold = 0.0;
    private boolean printVariance = false;

    AttributeSelectedClassifier reducedClassifier;
    CfsSubsetEval evaluator;
    BestFirst search;

    private double[] learnSVM( Instances data, double c, double gamma ) throws
Exception {
        double[] result = new double[4];
        Evaluation eval = new Evaluation(data);
        SMO scheme = new SMO();
        String[] options = weka.core.Utils.splitOptions("-C "+c+" -E 1.0 -G "+gamma+" -
A 250007 -L 0.0010 -P 1.0E-12 -N 0 -R -V -1 -W 1");
        scheme.setOptions( options );
        eval = new Evaluation(data);
        eval.crossValidateModel(scheme, data, 10, new Random(1));
        double[][] matrix = eval.confusionMatrix();
        for (int i = 0; i < matrix.length; i++) {
            for (int j = 0; j < matrix[0].length; j++) {
                result[i*2+j] = matrix[i][j];
            }
        }

        return result;
    }

    private double[] learnLibSVM( Instances data, double c, double gamma, int svm_type,
int kernel ) throws Exception {
        double[] result = new double[4];
        Evaluation eval = new Evaluation(data);
        boolean selection = true;

              WLSVM lib = new WLSVM();
        String[] ops = weka.core.Utils.splitOptions("-S "+svm_type+" -K "+kernel+" -G
"+gamma+" -C " + c + " -M 100 -Z 1");
        lib.setOptions( ops );

        if( selection ) {
            reducedClassifier.setClassifier(lib);
            reducedClassifier.setEvaluator(evaluator);
            reducedClassifier.setSearch(search);
            eval = new Evaluation(data);
            eval.crossValidateModel(reducedClassifier, data, 10, new Random(1));
        } else {
            eval = new Evaluation(data);
            eval.crossValidateModel(lib, data, 10, new Random(1));
        }
        // lib.buildClassifier(data);
        // eval.evaluateModel(lib, data);
        double[][] matrix = eval.confusionMatrix();
        for (int i = 0; i < matrix.length; i++) {
            for (int j = 0; j < matrix[0].length; j++) {
```

```
                result[i*2+j] = matrix[i][j];
            }
        }

        return result;
    }


    public SVMGrid() {}

    public SVMGrid( String file, int svm, int kernel,
                                int c_count, double c_start, double c_step,
                                int gamma_count, double gamma_start,
                                double gamma_step,
                                double variance_threshold, boolean printVariance ) {
        try {
            Instances data = new Instances(new BufferedReader(new FileReader(file)));
            data.setClassIndex( data.numAttributes() - 1);
            System.out.println( data.toSummaryString() );
            System.out.println( "Class index: " + data.classIndex() );
            System.out.println( "read: " + data.numInstances() + " instances" );

            this.svm = svm;
            this.kernel = kernel;
            this.c_start = c_start;
            this.c_count = c_count;
            this.c_step = c_step;
            this.gamma_start = gamma_start;
            this.gamma_count = gamma_count;
            this.gamma_step = gamma_step;
            this.variance_threshold = variance_threshold;
            this.printVariance = printVariance;

            /*
             * feature selection
             *
             * if variance_threshold is set, i.e. > 0.0, remove all attributes with
             * a variance below this limit
             */
            if (variance_threshold > 0) {
                System.out.println( "Removing attributes with variance < " +
variance_threshold );
                Vector lowVariance = new Vector(); // store indices of low variance
attributes
                for (int i = 0; i < data.numAttributes(); i++) {
                    if (data.attribute(i).isNumeric()) {
                        if( printVariance)
                            System.out.println(data.attribute(i).toString() + ":\t " +
data.variance(i));
                        if (data.variance(i) < variance_threshold) {
                            lowVariance.addElement(i);
                        }
                    }
                }

                // remove low variance attributes
                // because index changes upon removal, keep track of
                // number of changes and adjust index
                for (int i = 0; i < lowVariance.size(); i++) {
                    int index = (Integer) lowVariance.elementAt(i);
                    // System.out.println("Deleting index " + index + " (i=" + i + ") -
> new index: " + (index - i) + " size: " + data.numAttributes());
                    index = index - i;
                    data.deleteAttributeAt(index);
                }

                System.out.println("\nNumber of attributes after removal: " +
data.numAttributes() +"\n");
            }
```

```
            reducedClassifier = new AttributeSelectedClassifier();
            evaluator = new CfsSubsetEval();
            search = new BestFirst();

            results = new double[c_count*gamma_count][6];
            values = new double[c_count][gamma_count]; // plot values
            labels = new double[c_count][gamma_count]; // label values

            for( int i=0; i<c_count; i++ ) {
                double c = Math.pow( 2, c_start+c_step*i );
                labels[0][i] = c;
                for( int j=0; j<gamma_count; j++ ) {
                    double gamma = Math.pow( 2, gamma_start+gamma_step*j );
                    try {
                        double[] res = learnLibSVM( data, c, gamma, svm, kernel );
                        results[i*gamma_count+j][0] = c;
                        results[i*gamma_count+j][1] = gamma;
                        for( int k=0; k<res.length; k++ )
                            results[i*gamma_count+j][k+2] = res[k];
                        values[i][j] = (
(res[0]/(res[0]+res[1]))+(res[3]/(res[2]+res[3])) ) * 0.5; // CCR
                        labels[1][j] = gamma;
                        System.out.println( "run " + (i*gamma_count+j+1) + "/" +
(c_count*gamma_count) + ", c: " + c + ", gamma: " + gamma + ", CCR: " + values[i][j]);
                    } catch( Exception e ) {
                        e.printStackTrace();
                    }

                }
            }
        } catch( IOException ioe ) {
            ioe.printStackTrace();
        }
    }

    public static void main(String[] args) {
    }

    /**
     * @return the values
     */
    public double[][] getValues() {
        return values;
    }

    /**
     * @return the labels
     */
    public double[][] getLabels() {
        return labels;
    }

    /**
     * @return the results
     */
    public double[][] getResults() {
        return results;
    }

    /**
     * @return the c_count
     */
    public int getC_count() {
        return c_count;
    }

    /**
     * @param c_count the c_count to set
```

```
     */
    public void setC_count(int c_count) {
        this.c_count = c_count;
    }

    /**
     * @return the c_start
     */
    public double getC_start() {
        return c_start;
    }

    /**
     * @param c_start the c_start to set
     */
    public void setC_start(double c_start) {
        this.c_start = c_start;
    }

    /**
     * @return the c_step
     */
    public double getC_step() {
        return c_step;
    }

    /**
     * @param c_step the c_step to set
     */
    public void setC_step(double c_step) {
        this.c_step = c_step;
    }

    /**
     * @return the gamma_count
     */
    public int getGamma_count() {
        return gamma_count;
    }

    /**
     * @param gamma_count the gamma_count to set
     */
    public void setGamma_count(int gamma_count) {
        this.gamma_count = gamma_count;
    }

    /**
     * @return the gamma_start
     */
    public double getGamma_start() {
        return gamma_start;
    }

    /**
     * @param gamma_start the gamma_start to set
     */
    public void setGamma_start(double gamma_start) {
        this.gamma_start = gamma_start;
    }

    /**
     * @return the gamma_step
     */
    public double getGamma_step() {
        return gamma_step;
    }

    /**
```

```
     * @param gamma_step the gamma_step to set
     */
    public void setGamma_step(double gamma_step) {
        this.gamma_step = gamma_step;
    }

    /**
     * @return the variance_threshold
     */
    public double getVariance_threshold() {
        return variance_threshold;
    }

    /**
     * @param variance_threshold the variance_threshold to set
     */
    public void setVariance_threshold(double variance_threshold) {
        this.variance_threshold = variance_threshold;
    }

    /**
     * @return the printVariance
     */
    public boolean isPrintVariance() {
        return printVariance;
    }

    /**
     * @param printVariance the printVariance to set
     */
    public void setPrintVariance(boolean printVariance) {
        this.printVariance = printVariance;
    }

    /**
     * @return the svm
     */
    public int getSvm() {
        return svm;
    }

    /**
     * @param svm the svm to set
     */
    public void setSvm(int svm) {
        this.svm = svm;
    }

    /**
     * @return the kernel
     */
    public int getKernel() {
        return kernel;
    }

    /**
     * @param kernel the kernel to set
     */
    public void setKernel(int kernel) {
        this.kernel = kernel;
    }

}
```

7.2. Bibliography

1. Geddes, A., *80th Anniversary of the discovery of penicillin An appreciation of Sir Alexander Fleming.* Int J Antimicrob Agents, 2008.
2. Djerassi, C., *The mother of the pill.* Recent Prog Horm Res, 1995. **50**: p. 1-17.
3. Hofmann, A., *LSD - mein Sorgenkind. Die Entdeckung einer "Wunderdroge"*. 1993, Munich: Dtv.
4. Mahdi, J.G., et al., *The historical analysis of aspirin discovery, its relation to the willow tree and antiproliferative and anticancer potential.* Cell Prolif, 2006. **39**(2): p. 147-55.
5. Lahana, R., *How many leads from HTS?* Drug Discov Today, 1999. **4**(10): p. 447-448.
6. Johnson, M.A. and G.M. Maggiora, *Concepts and Applications of Molecular Similarity*. 1990, New York: John Wiley.
7. Kopp, H., *Über die Vorausbestimmung des specifischen Gewichts einiger Klassen chemischer Verbindungen.* Annalen der Physik, 1839. **123**(5): p. 133-153.
8. Kopp, H., *Über die Vorausbestimmung einiger physikalischen Eigenschaften bei mehreren Reihen organischer Verbindungen.* Ann. Chem., 1842. **41**(1): p. 79-89.
9. Kopp, H., *Über die Vorausbestimmung einiger physikalischen Eigenschaften bei mehreren Reihen organischer Verbindungen.* Ann. Chem., 1842. **41**(2): p. 162-168.
10. Bonchev, D. and D.H. Rouvray, *Chemical Graph Theory: Introduction and Fundamentals*. Mathematical Chemistry Series, ed. D.H. Rouvray. Vol. 1. 1991, New York: Abacus.
11. Crum-Brown, A. and T. Fraser, *On the connection between chemical constitution and physiological action.* Trans. Roy. Soc. Edinburgh, 1868-1869. **25**: p. 1-53.
12. Overton, E., *Studien über die Narkose, zugleich ein Beitrag zur Allgemeinen Pharmakologie*. 1901, Gustav Fischer: Jena, Germany. p. 1-195.
13. Meyer, H., *Zur Theorie der Alkoholnarkose.* Arch Exp Pathol Pharmakol, 1899. **42**: p. 109-118.
14. Meyer, K.H. and H. Gottlieb-Billroth, *Theorie der Narkose durch Inhalationsanästhetika.* Z Physiol Chem, 1920. **112**: p. 55-79.
15. Dearden, J.C., *Partioning and Lipophilicity in Quantitative Structure-Activity Relationssships.* Environmental Health Perspectives, 1985. **61**: p. 203-228.
16. Lynch, C., *Meyer and Overton revisited.* Anesthesia and Analgesia, 2008. **107**(3): p. 864-867.
17. Hammett, L.P., *The Effect of Structure upon the Reactions of Organic Compounds.* J. Am. Chem. Soc., 1937. **59**(1): p. 96-103.
18. Hansch, C., et al., *Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients.* Nature, 1962. **194**(4824): p. 178-180.
19. Taft, R.W., *Steric effects in organic chemistry*, M.S. Newman, Editor. 1956, Wiley: New York. p. 556.
20. Kier, L.B. and L.H. Hall, *Molecular Structure Description: The Electrotopological State*. 1999, San Diego, CA: Academic Press.
21. Livingstone, D., *Data Analysis for Chemists: Applications to QSAR and Chemical Product Design*. 1st ed. 1995, Oxford: Oxford University Press.
22. Free, S.M. and J.W. Wilson, *A Mathematical Contribution to Structure-Activity Studies.* Journal of Medicinal Chemistry, 1964. **7**(4): p. 395-399.
23. Schultz, T.W., et al., *Quantitative structure-activity relationships (QSARs) in toxicology: a historical perspective.* Journal of Molecular Structure: THEOCHEM, 2003. **622**: p. 1-22.
24. Dong, A., J. Wei, and Q. Gao, *3D-pharmacophore model for RXR(gamma) agonists.* Neurochem Int, 2008.
25. Greener, M. (2005) *QSAR: Prediction Beyond the Fourth Dimension*. Drug Discovery & Development.
26. Lipinski, C.A., et al., *Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings.* Adv. Drug Del. Rev., 1997. **23**: p. 3-25.
27. Veber, D.F., et al., *Molecular properties that influence the oral bioavailability of drug candidates.* J. Med. Chem, 2002. **45**(12): p. 2615-2623.
28. Böhm, H.-J. and G. Schneider, eds. *Virtual Screening for Bioactive Molecules*. Methods and Principles in Medicinal Chemistry, ed. R. Mannhold, H. Kubinyi, and H. Timmerman. 2000, Wiley-VCH: Weinheim, New York.
29. Russel, S. and P. Norvig, *Artificial Intelligence: A Modern Approach*. 2nd ed. 2002, New Jersey: Prentice Hall.
30. Turing, A.M., *Computing Machinery and Intelligence.* Mind, 1950. **LIX**(236): p. 433-460.
31. Theodoridis, S. and K. Koutroumbas, *Pattern Recognition*. 3rd ed. 2006: Academic Press.
32. Grahame-Smith, D.G. and J.K. Aronson, *Oxford Textbook of Clinical Pharmacology and Drug Therapy*. 3rd ed. 2002, Oxford: Oxford University Press.

33. Xu, C., C.Y. Li, and A.N. Kong, *Induction of phase I, II and III drug metabolism/transport by xenobiotics.* Arch Pharm Res, 2005. **28**(3): p. 249-68.
34. Tukey, R.H. and C.P. Strassburg, *Human UDP-glucuronosyltransferases: metabolism, expression, and disease.* Annu Rev Pharmacol Toxicol, 2000. **40**: p. 581-616.
35. Aono, S., et al., *Analysis of genes for bilirubin UDP-glucuronosyltransferase in Gilbert's syndrome.* Lancet, 1995. **345**(8955): p. 958-9.
36. Aono, S., et al., *A new type of defect in the gene for bilirubin uridine 5'-diphosphate-glucuronosyltransferase in a patient with Crigler-Najjar syndrome type I.* Pediatr Res, 1994. **35**(6): p. 629-32.
37. Guillemette, C., *Pharmacogenomics of human UDP-glucuronosyltransferase enzymes.* Pharmacogenomics J, 2003. **3**(3): p. 136-58.
38. Schwarz, U.I., *Clinical relevance of genetic polymorphisms in the human CYP2C9 gene.* Eur J Clin Invest, 2003. **33 Suppl 2**: p. 23-30.
39. Zanger, U.M., et al., *Functional pharmacogenetics/genomics of human cytochromes P450 involved in drug biotransformation.* Anal Bioanal Chem, 2008. **392**(6): p. 1093-108.
40. Wojnowski, L. and L.K. Kamdem, *Clinical implications of CYP3A polymorphisms.* Expert Opin Drug Metab Toxicol, 2006. **2**(2): p. 171-82.
41. Droll, K., et al., *Comparison of three CYP2D6 probe substrates and genotype in Ghanaians, Chinese and Caucasians.* Pharmacogenetics, 1998. **8**(4): p. 325-33.
42. Conney, A.H., *Pharmacological implications of microsomal enzyme induction.* Pharmacol Rev, 1967. **19**(3): p. 317-66.
43. Schoedel, K.A. and R.F. Tyndale, *Induction of nicotine-metabolizing CYP2B1 by ethanol and ethanol-metabolizing CYP2E1 by nicotine: summary and implications.* Biochim Biophys Acta, 2003. **1619**(3): p. 283-90.
44. Graham, M.J. and B.G. Lake, *Induction of drug metabolism: species differences and toxicological relevance.* Toxicology, 2008. **254**(3): p. 184-91.
45. Murray, M., *Altered CYP expression and function in response to dietary factors: potential roles in disease pathogenesis.* Curr Drug Metab, 2006. **7**(1): p. 67-81.
46. Babu, P.V. and D. Liu, *Green tea catechins and cardiovascular health: an update.* Curr Med Chem, 2008. **15**(18): p. 1840-50.
47. Zhou, S.F., et al., *Clinically important drug interactions potentially involving mechanism-based inhibition of cytochrome P450 3A4 and the role of therapeutic drug monitoring.* Ther Drug Monit, 2007. **29**(6): p. 687-710.
48. Foti, R.S. and J.L. Wahlstrom, *Prediction of CYP-mediated drug interactions in vivo using in vitro data.* IDrugs, 2008. **11**(12): p. 900-5.
49. Lin, J.H., *CYP induction-mediated drug interactions: in vitro assessment and clinical implications.* Pharm Res, 2006. **23**(6): p. 1089-116.
50. Evans, W.E. and M.V. Relling, *Pharmacogenomics: translating functional genomics into rational therapeutics.* Science, 1999. **286**(5439): p. 487-91.
51. Kiang, T.K., M.H. Ensom, and T.K. Chang, *UDP-glucuronosyltransferases and clinical drug-drug interactions.* Pharmacol Ther, 2005. **106**(1): p. 97-132.
52. Sugatani, J., et al., *The induction of human UDP-glucuronosyltransferase 1A1 mediated through a distal enhancer module by flavonoids and xenobiotics.* Biochem Pharmacol, 2004. **67**(5): p. 989-1000.
53. Carlson, S.D., et al., *Blood barriers of the insect.* Annu Rev Entomol, 2000. **45**: p. 151-74.
54. Cserr, H.F. and M. Bundgaard, *Blood-brain interfaces in vertebrates: a comparative approach.* Am J Physiol, 1984. **246**(3 Pt 2): p. R277-88.
55. Ehrlich, P., *Das Sauerstoff-Bedürfniss des Organismus: eine farbenanalytische Studie*. 1885, Verlag von August Hirschwald. p. 1-167.
56. Goldmann, E.E., *Vitalfärbung am Zentralnervensystem. Beitrag zur Physio-pathologie des Plexus Chorioideus und der Hirnhäute*. 1913.
57. Spatz, H., *Die Bedeutung der vitalen Färbung für die Lehre vom Stoffaustausch zwischen dem Zentralnervensystem und dem übrigen Körper.* Archiv für Psychiatrie und Nervenkrankheiten, 1934. **101**(1).
58. Reese, T.S. and M.J. Karnovsky, *Fine structural localization of a blood-brain barrier to exogenous peroxidase.* J Cell Biol, 1967. **34**(1): p. 207-17.
59. Brightman, M.W. and T.S. Reese, *Junctions between intimately apposed cell membranes in the vertebrate brain.* J Cell Biol, 1969. **40**(3): p. 648-77.

60. Krogh, A., *The active and passive exchanges of inorganic ions through the surfaces of living cells and through living membranes generally.* Proceedings of the Royal Society of London, 1946. **133**(871): p. 140-200.
61. Pardridge, W.M., *Blood-brain barrier genomics and the use of endogenous transporters to cause drug penetration into the brain.* Curr Opin Drug Discov Devel, 2003. **6**(5): p. 683-91.
62. Pardridge, W.M., *Blood-brain barrier drug targeting: the future of brain drug development.* Mol Interv, 2003. **3**(2): p. 90-105, 51.
63. Rodriguez-Baeza, A., et al., *Morphological features in human cortical brain microvessels after head injury: a three-dimensional and immunocytochemical study.* Anat Rec A Discov Mol Cell Evol Biol, 2003. **273**(1): p. 583-93.
64. Abbott, N.J., *Dynamics of CNS barriers: evolution, differentiation, and modulation.* Cell Mol Neurobiol, 2005. **25**(1): p. 5-23.
65. Liu, X., C. Chen, and B.J. Smith, *Progress in brain penetration evaluation in drug discovery and development.* J Pharmacol Exp Ther, 2008. **325**(2): p. 349-56.
66. Sims, D.E., *Recent advances in pericyte biology--implications for health and disease.* Can J Cardiol, 1991. **7**(10): p. 431-43.
67. Balabanov, R., et al., *CNS microvascular pericytes express macrophage-like function, cell surface integrin alpha M, and macrophage marker ED-2.* Microvasc Res, 1996. **52**(2): p. 127-42.
68. Pardridge, W.M., *Blood-brain barrier genomics.* Stroke, 2007. **38**(2 Suppl): p. 686-90.
69. Pardridge, W.M., *The blood-brain barrier: bottleneck in brain drug development.* NeuroRx, 2005. **2**(1): p. 3-14.
70. Omidi, Y., et al., *Evaluation of the immortalised mouse brain capillary endothelial cell line, b.End3, as an in vitro blood-brain barrier model for drug uptake and transport studies.* Brain Res, 2003. **990**(1-2): p. 95-112.
71. Zhang, Y., et al., *Plasma membrane localization of multidrug resistance-associated protein homologs in brain capillary endothelial cells.* J Pharmacol Exp Ther, 2004. **311**(2): p. 449-55.
72. Gao, B. and P.J. Meier, *Organic anion transport across the choroid plexus.* Microsc Res Tech, 2001. **52**(1): p. 60-4.
73. Cordon-Cardo, C., et al., *Expression of the multidrug resistance gene product (P-glycoprotein) in human normal and tumor tissues.* J Histochem Cytochem, 1990. **38**(9): p. 1277-87.
74. Chin, J.E., et al., *Structure and expression of the human MDR (P-glycoprotein) gene family.* Mol Cell Biol, 1989. **9**(9): p. 3808-20.
75. Smit, J.J., et al., *Homozygous disruption of the murine mdr2 P-glycoprotein gene leads to a complete absence of phospholipid from bile and to liver disease.* Cell, 1993. **75**(3): p. 451-62.
76. Hsu, S.I., et al., *Structural analysis of the mouse mdr1a (P-glycoprotein) promoter reveals the basis for differential transcript heterogeneity in multidrug-resistant J774.2 cells.* Mol Cell Biol, 1990. **10**(7): p. 3596-606.
77. Annese, V., et al., *Multidrug resistance 1 gene in inflammatory bowel disease: a meta-analysis.* World J Gastroenterol, 2006. **12**(23): p. 3636-44.
78. Jones, P.M. and A.M. George, *Symmetry and structure in P-glycoprotein and ABC transporters what goes around comes around.* Eur J Biochem, 2000. **267**(17): p. 5298-305.
79. Saurin, W., M. Hofnung, and E. Dassa, *Getting in or out: early segregation between importers and exporters in the evolution of ATP-binding cassette (ABC) transporters.* J Mol Evol, 1999. **48**(1): p. 22-41.
80. McDevitt, C.A. and R. Callaghan, *How can we best use structural information on P-glycoprotein to design inhibitors?* Pharmacol Ther, 2007. **113**(2): p. 429-41.
81. Kerb, R., S. Hoffmeyer, and U. Brinkmann, *ABC drug transporters: hereditary polymorphisms and pharmacological impact in MDR1, MRP1 and MRP2.* Pharmacogenomics, 2001. **2**(1): p. 51-64.
82. Sparreboom, A., et al., *Pharmacogenomics of ABC transporters and its role in cancer chemotherapy.* Drug Resist Updat, 2003. **6**(2): p. 71-84.
83. Wang, D. and W. Sadee, *Searching for polymorphisms that affect gene expression and mRNA processing: example ABCB1 (MDR1).* Aaps J, 2006. **8**(3): p. E515-20.
84. Komoto, C., et al., *MDR1 haplotype frequencies in Japanese and Caucasian, and in Japanese patients with colorectal cancer and esophageal cancer.* Drug Metab Pharmacokinet, 2006. **21**(2): p. 126-32.
85. Hitzl, M., et al., *The C3435T mutation in the human MDR1 gene is associated with altered efflux of the P-glycoprotein substrate rhodamine 123 from CD56+ natural killer cells.* Pharmacogenetics, 2001. **11**(4): p. 293-8.
86. Petraccia, L., et al., *MDR (multidrug resistance) in hepatocarcinoma clinical-therapeutic implications.* Clin Ter, 2003. **154**(5): p. 325-35.

87. Zhou, S.F., *Structure, function and regulation of P-glycoprotein and its clinical relevance in drug disposition.* Xenobiotica, 2008. **38**(7-8): p. 802-32.
88. Mizutani, T., et al., *Genuine functions of P-glycoprotein (ABCB1).* Curr Drug Metab, 2008. **9**(2): p. 167-74.
89. Fiocchi, C., *Inflammatory bowel disease: etiology and pathogenesis.* Gastroenterology, 1998. **115**(1): p. 182-205.
90. Yacyshyn, B., W. Maksymowych, and M.B. Bowen-Yacyshyn, *Differences in P-glycoprotein-170 expression and activity between Crohn's disease and ulcerative colitis.* Hum Immunol, 1999. **60**(8): p. 677-87.
91. Rescigno, M., *The pathogenic role of intestinal flora in IBD and colon cancer.* Curr Drug Targets, 2008. **9**(5): p. 395-403.
92. Henckaerts, L., et al., *The role of genetics in inflammatory bowel disease.* Curr Drug Targets, 2008. **9**(5): p. 361-8.
93. Gutmann, H., et al., *Breast cancer resistance protein and P-glycoprotein expression in patients with newly diagnosed and therapy-refractory ulcerative colitis compared with healthy controls.* Digestion, 2008. **78**(2-3): p. 154-62.
94. Englund, G., et al., *Efflux transporters in ulcerative colitis: decreased expression of BCRP (ABCG2) and Pgp (ABCB1).* Inflamm Bowel Dis, 2007. **13**(3): p. 291-7.
95. Buchman, A.L., et al., *A higher dose requirement of tacrolimus in active Crohn's disease may be related to a high intestinal P-glycoprotein content.* Dig Dis Sci, 2005. **50**(12): p. 2312-5.
96. Sambuelli, A.M., et al., *Multidrug resistance gene (MDR-1) expression in the colonic mucosa of patients with refractory ulcerative colitis.* Acta Gastroenterol Latinoam, 2006. **36**(1): p. 23-32.
97. Hughes, J.R., *One of the hottest topics in epileptology: ABC proteins. Their inhibition may be the future for patients with intractable seizures.* Neurol Res, 2008. **30**(9): p. 920-5.
98. Stouch, T.R. and O. Gudmundsson, *Progress in understanding the structure-activity relationships of P-glycoprotein.* Adv Drug Deliv Rev, 2002. **54**(3): p. 315-28.
99. Tsuruo, T., et al., *Overcoming of vincristine resistance in P388 leukemia in vivo and in vitro through enhanced cytotoxicity of vincristine and vinblastine by verapamil.* Cancer Res, 1981. **41**(5): p. 1967-72.
100. Coley, H.M., P.R. Twentyman, and P. Workman, *Improved cellular accumulation is characteristic of anthracyclines which retain high activity in multidrug resistant cell lines, alone or in combination with verapamil or cyclosporin A.* Biochem Pharmacol, 1989. **38**(24): p. 4467-75.
101. Chambers, S.K., et al., *Enhancement of anthracycline growth inhibition in parent and multidrug-resistant Chinese hamster ovary cells by cyclosporin A and its analogues.* Cancer Res, 1989. **49**(22): p. 6275-9.
102. Hait, W.N., et al., *Terfenadine (Seldane): a new drug for restoring sensitivity to multidrug resistant cancer cells.* Biochem Pharmacol, 1993. **45**(2): p. 401-6.
103. Germann, U.A., et al., *Chemosensitization and drug accumulation effects of VX-710, verapamil, cyclosporin A, MS-209 and GF120918 in multidrug resistant HL60/ADR cells expressing the multidrug resistance-associated protein MRP.* Anticancer Drugs, 1997. **8**(2): p. 141-55.
104. Twentyman, P.R. and N.M. Bleehen, *Resistance modification by PSC-833, a novel non-immunosuppressive cyclosporin [corrected].* Eur J Cancer, 1991. **27**(12): p. 1639-42.
105. Kim, R.B., et al., *Interrelationship between substrates and inhibitors of human CYP3A and P-glycoprotein.* Pharm Res, 1999. **16**(3): p. 408-14.
106. Benet, L.Z. and C.L. Cummins, *The drug efflux-metabolism alliance: biochemical aspects.* Adv Drug Deliv Rev, 2001. **50 Suppl 1**: p. S3-11.
107. Hyafil, F., et al., *In vitro and in vivo reversal of multidrug resistance by GF120918, an acridonecarboxamide derivative.* Cancer Res, 1993. **53**(19): p. 4595-602.
108. Slate, D.L., et al., *RS-33295-198: a novel, potent modulator of P-glycoprotein-mediated multidrug resistance.* Anticancer Res, 1995. **15**(3): p. 811-4.
109. Roe, M., et al., *Reversal of P-glycoprotein mediated multidrug resistance by novel anthranilamide derivatives.* Bioorg Med Chem Lett, 1999. **9**(4): p. 595-600.
110. Newman, M.J., et al., *Discovery and characterization of OC144-093, a novel inhibitor of P-glycoprotein-mediated multidrug resistance.* Cancer Res, 2000. **60**(11): p. 2964-72.
111. Wang, R.B., et al., *Structure-activity relationship: analyses of p-glycoprotein substrates and inhibitors.* J Clin Pharm Ther, 2003. **28**(3): p. 203-28.
112. Hait, N.C., et al., *Role of sphingosine kinase 2 in cell migration toward epidermal growth factor.* J Biol Chem, 2005. **280**(33): p. 29462-9.
113. Owen, A., B. Chandler, and D.J. Back, *The implications of P-glycoprotein in HIV: friend or foe?* Fundam Clin Pharmacol, 2005. **19**(3): p. 283-96.

114. Martin, C., et al., *Communication between multiple drug binding sites on P-glycoprotein.* Mol Pharmacol, 2000. **58**(3): p. 624-32.
115. Zamora, J.M., H.L. Pearce, and W.T. Beck, *Physical-chemical properties shared by compounds that modulate multidrug resistance in human leukemic cells.* Mol Pharmacol, 1988. **33**(4): p. 454-62.
116. Ueda, K., Y. Taguchi, and M. Morishima, *How does P-glycoprotein recognize its substrates?* Semin Cancer Biol, 1997. **8**(3): p. 151-9.
117. Seelig, A. and E. Landwojtowicz, *Structure-activity relationship of P-glycoprotein substrates and modifiers.* Eur J Pharm Sci, 2000. **12**(1): p. 31-40.
118. Koshiba, S., et al., *Human ABC transporters ABCG2 (BCRP) and ABCG4.* Xenobiotica, 2008. **38**(7-8): p. 863-88.
119. Jonker, J.W., et al., *Role of breast cancer resistance protein in the bioavailability and fetal penetration of topotecan.* J Natl Cancer Inst, 2000. **92**(20): p. 1651-6.
120. Maliepaard, M., et al., *Overexpression of the BCRP/MXR/ABCP gene in a topotecan-selected ovarian tumor cell line.* Cancer Res, 1999. **59**(18): p. 4559-63.
121. Allen, J.D., et al., *The mouse Bcrp1/Mxr/Abcp gene: amplification and overexpression in cell lines selected for resistance to topotecan, mitoxantrone, or doxorubicin.* Cancer Res, 1999. **59**(17): p. 4237-41.
122. Wigner, E., *The Unreasonable Effectiveness of Mathematics in the Natural Sciences.* Communications on Pure and Applied Mathematics, 1960. **13**(1): p. 1-14.
123. Sylvester, J.J., *Chemistry and Algebra.* Nature, 1878. **17**(432): p. 284.
124. Chartrand, G. and P. Zhang, *Introduction to Graph Theory*. 1st ed. The Walter Rudin Student Series in Advanced Mathematics. 2004, Columbus, Ohio: McGraw-Hill.
125. Weininger, D., *SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules.* J Chem Inf Comput Sci, 1988. **28**(1): p. 31-36.
126. Weininger, D., A. Weininger, and J.L. Weininger, *SMILES. 2. Algorithm for generation of unique SMILES notation.* J Chem Inf Comput Sci, 1989. **29**(2): p. 97-101.
127. Daylight Chemical Information Systems, I., *Daylight Theory Manual*. 2008, Daylight Chemical Information Systems, Inc.
128. Barnard, J.M., *Substructure Searching Methods: Old and New.* Journal of Chemical Informatics and Computational Sciences, 1993. **33**(4): p. 532-538.
129. Karp, R.M., *On the computational complexity of combinatorial problems.* Networks, 1975. **5**: p. 45-68.
130. Willett, P., J.M. Barnard, and G.M. Downs, *Chemical Similarity Searching.* J. Chem. Inf. Comput. Sci, 1998. **38**: p. 983-996.
131. Willett, P., *Similarity-based virtual screening using 2D fingerprints.* Drug Discov Today, 2006. **11**(23-24): p. 1046-53.
132. Hubálek, Z., *Coefficients of Association and Similarity, Based on Binary (Presence-Absence) Data: an Evaluation.* Biol Rev Cambridge Philos Soc, 1982. **1982**(57): p. 669-689.
133. Perez, J.J., *Managing molecular diversity.* Chem Soc Rev, 2005. **34**: p. 143-152.
134. Leach, A.R., *Molecular Modelling - Principles and Applications*. 2nd ed. 2001: Prentice Hall.
135. Petitjean, M., *Applications of the Radius-Diameter Diagram to the Classification of Topological and Geometrical Shapes of Chemical Compounds.* J Chem Inf Comput Sci, 1992. **32**: p. 331-337.
136. Nelson, D.L. and M.M. Cox, *Lehninger Principles of Biochemistry*. 4th ed. 2004: W. H. Freeman.
137. Cleland, W.W., *The use of isotope effects to determine enzyme mechanisms.* Arch Biochem Biophys, 2005. **433**(1): p. 2-12.
138. Todeschini, R. and V. Consonni, eds. *Handbook of Molecular Descriptors*. Methods and Principles in Medicinal Chemistry, ed. R. Mannhold, H. Kubinyi, and H. Timmerman. 2000, Wiley-VCH: Weinheim, New York.
139. Housecroft, C. and E. Constable, *Chemistry*. 3rd ed. 2006: Prentice Hall.
140. Gutman, I., *Topological Formulas for Free-Valency Index.* Croat. Chem. Acta, 1978. **51**: p. 29-33.
141. Grüber, C. and V. Buss, *Quantum-Mechanically Calculated Properties for the Development of Quantitative Structure-Activity Relationships (QSARs). pKa-Values of Phenols and Aromatic and Aliphatic Carboxylic Acids.* Chemosphere, 1989. **19**: p. 1595-1609.
142. Rouvray, D.H., *Do Molecular Models Accurately Reflect Reality?* Chemist in Industry, 1997. **15**: p. 587-590.
143. Rouvray, D.H., *Atoms as Hard Spheres.* Chemistry in Britain, 2000. **36**.
144. Stanton, D. and P. Jurs, *Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure-property relationship studies.* Analytical Chemistry, 1990. **62**(21): p. 2323-2329.

145. Lee, B. and F.M. Richards, *The interpretation of protein structures: estimation of static accessibility.* J Mol Biol, 1971. **55**(3): p. 379-400.
146. Ertl, P., B. Rohde, and P. Selzer, *Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties.* J Med Chem, 2000. **43**(20): p. 3714-7.
147. Winiwarter, S., et al., *Correlation of human jejunal permeability (in vivo) of drugs with experimentally and theoretically derived parameters. A multivariate data analysis approach.* J Med Chem, 1998. **41**(25): p. 4939-49.
148. Palm, K., et al., *Polar molecular surface properties predict the intestinal absorption of drugs in humans.* Pharm Res, 1997. **14**(5): p. 568-71.
149. Palm, K., et al., *Correlation of drug absorption with molecular surface properties.* J Pharm Sci, 1996. **85**(1): p. 32-9.
150. Clark, D.E., *Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 2. Prediction of blood-brain barrier penetration.* J Pharm Sci, 1999. **88**(8): p. 815-21.
151. Clark, D.E., *Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 1. Prediction of intestinal absorption.* J Pharm Sci, 1999. **88**(8): p. 807-14.
152. Katritzky, A.R. and E.V. Gordeeva, *Traditional topological indices vs electronic, geometrical, and combined molecular descriptors in QSAR/QSPR research.* J Chem Inf Comput Sci, 1993. **33**(6): p. 835-57.
153. Barker, J.L., *Activity of CNS depressants related to hydrophobicity.* Nature, 1974. **252**(5478): p. 52-54.
154. Fujita, T., J. Iwasa, and C. Hansch, *A New Substituent Constant, π, Derived from Partition Coefficients.* Journal of the American Chemical Society, 1964. **86**(23): p. 5175-5180.
155. Leo, A.J., *Calculating log Poct from Structures.* Chemical Reviews, 1993. **93**(4): p. 1281-1306.
156. Ghose, A.K. and G.M. Crippen, *Atomic Physicochemical Parameters for Three-Dimensional Structure-Directed Quantitative Structure-Activity Relationships I. Partition Coefficients as a Measure of Hydrophobicity.* Journal of Computational Chemistry, 1986. **7**(4).
157. Wang, R., Y. Fu, and L. Lai, *A New Atom-Additive Method for Calculating Parition Coefficients.* Journal of Chemical Informatics and Computational Sciences, 1997. **37**: p. 615-521.
158. Klopman, G. and S. Wang, *A computer automated structure evaluation (CASE) approach to calculation of partition coefficient.* J Comput Chem, 1991. **12**: p. 1025-1032.
159. Viswanadhan, V.N., A.K. Ghose, and J.J. Wendoloski, *Estimating aqueous solvation and lipophilicity of small organic molecules: A comparative overview of atom/group contribution methods* Perspectives in Drug Discovery and Design, 2000. **9**(1): p. 85-98.
160. Kerns, E.H. and L. Di, *Drug-like properties: concepts, structure design and methods: from ADME to toxicity optimization*. 2008: Academic Pr.
161. Wiener, H., *Structural Determination of Paraffin Boiling Points.* Journal of the American Chemical Society, 1947. **69**(1): p. 17-20.
162. Platt, J.R., *Influence of Neighbor Bonds on Additive Bond Properties in Paraffins.* Journal of Chemical Physics, 1947. **15**(6): p. 419-420.
163. Randic, M., *Characterization of molecular branching.* Journal of the American Chemical Society, 1975. **97**(23): p. 6609-6615.
164. Kier, L.B. and L.H. Hall, *Intermolecular accessibility: the meaning of molecular connectivity.* J Chem Inf Comput Sci, 2000. **40**(3): p. 792-5.
165. Kier, L.B. and L.H. Hall, *Molecular Connectivity in Chemistry and Drug Research*. 1976, New York/London: Academic Press.
166. Kier, L.B. and L.H. Hall, *General definition of valence delta-values for molecular connectivity.* J Pharm Sci, 1983. **72**(10): p. 1170-3.
167. Kier, L.B., *A Shape Index from Molecular Graphs.* Quantitative Structure-Activity Relationships, 1985. **4**(3): p. 109-116.
168. Kier, L.B. and L.H. Hall, *An electrotopological-state index for atoms in molecules.* Pharm Res, 1990. **7**(8): p. 801-7.
169. Gutman, I., *A Formula for the Wiener Number of Trees and its Extension to Graphs Containing Cycles.* Graph Theory Notes New York, 1994. **27**: p. 9-15.
170. Gutman, I. and S. Klavzar, *A method for calculating Wiener numbers of benzenoid hydrocarbons.* Models Chem., 1996. **133**(4): p. 389-399.
171. Bonchev, D., *The overall Wiener index - a new tool for characterization of molecular topology.* J Chem Inf Comput Sci, 2001. **41**(3): p. 582-92.

172. Tanford, C., *Physical Chemistry of Macromolecules, 1961*. 1961: John Wiley & Sons, Inc.
173. Rissanen, J., *Modeling by shortest data description.* Automatica, 1978. **14**(5): p. 465-471.
174. Burgin, M., *Generalized Kolmogorov complexity and duality in theory of computations.* Notices of the Russian Academy of Sciences, 1982. **25**(3): p. 19-23.
175. Cover, T.M., P. Gacs, and R.M. Gray, *Kolmogorov's contributions to information theory and algorithmic complexity.* Annals of Probability, 1989. **17**(3): p. 840-865.
176. Lisboa, P.J.G. and S.J. Perantonis, *Complete solution of the local minima in the XOR problem.* Network: Computation in Neural Systems, 1991. **2**(1): p. 119-124.
177. Witten, I.H. and E. Frank, *Data mining: practical machine learning tools and techniques*. 2nd ed. Morgan Kaufmann series in data management systems, ed. J. Gray. 2005, San Francisco: Morgan Kaufmann.
178. Matthews, B.W., *Comparison of the predicted and observed secondary structure of T4 phage lysozyme.* Biochim Biophys Acta, 1975. **405**(2): p. 442-51.
179. Cohen, J., *A coefficient of agreement for nominal scales.* Educational and Psychological Measurement, 1960. **20**(1): p. 37-46.
180. Fleiss, J.L., *Measuring nominal scale agreement among many raters.* Psychological Bulletin, 1971. **76**(5): p. 378-382.
181. Landis, J.R. and G.G. Koch, *The measurement of observer agreement for categorical data.* Biometrics, 1977. **33**(1): p. 159-74.
182. Sim, J. and C.C. Wright, *The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements.* Physical Therapy, 2005. **85**(3): p. 257-268.
183. Cureton, E.E., *Validity, Reliability and Baloney*. Educational and Psychological Measurement, 1950. **10**: p. 94-96.
184. Mosier, C.I., *Symposium: The Need and Means of Cross-Validation. I. Problems and Designs of Cross-Validation.* Educational and Psychological Measurement, 1951. **11**(1): p. 5-11.
185. Kohavi, R., *A study of cross-validation and bootstrap for accuracy estimation and model selection.* Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, 1995. **2**(12): p. 1137-1143.
186. Kearns, M. and D. Ron, *Algorithmic stability and sanity-check bounds for leave-one-out cross-validation.* Neural Comput, 1999. **11**(6): p. 1427-53.
187. Rogers, W.H. and T.J. Wagner, *A finite sample distribution-free perfomance bound for local discrimination rules.* Annals of Statistics, 1978. **6**(3): p. 506-514.
188. Pomberger, G. and H. Dobler, *Algorithmen und Datenstrukturen*. 1st ed. 2008, München: Pearson.
189. Shannon, C.E., *A Mathematical Theory of Communication.* Bell System Technical Journal, 1948. **27**: p. 379-423.
190. Gini, C., *Variabilità e mutabilità.* Memorie di metodologica statistica, 1912.
191. Quinlan, J.R., *Induction of Decision Trees.* Machine Learning, 1986. **1**(1): p. 81-106.
192. Quinlan, J.R., *C4.5 Programs for Machine Learning*. 1993, Morgan Kaufmann.
193. Breiman, L., *Classification and Regression Trees*. 1984: Chapman & Hall / CRC.
194. Sonquist, J.A. and J.N. Morgan, *The Detection of Interaction Effects*. 1964: Survey Research Center, Institute for Social Research, University of Michigan. 296.
195. de Cerqueira Lima, P., et al., *Combinatorial QSAR modeling of P-glycoprotein substrates.* J Chem Inf Model, 2006. **46**(3): p. 1245-54.
196. Breiman, L., *Bagging Predictors.* Machine Learning, 1996. **24**(2): p. 123-140.
197. Davison, A.C. and D. Hinkley, *Bootstrap Methods and their Application*. Cambridge Series in Statistical and Probabilistic Mathematics. 2006: Cambridge University Press.
198. Breiman, L., *Random Forests.* Machine Learning, 2001. **45**(1): p. 5-32.
199. McCulloch, W. and W. Pitts, *A Logical Calculus of Ideas Immanent in Nervous Activity.* Bulletin of Mathematical Biophysics, 1943. **5**: p. 115-133.
200. Rosenblatt, F., *The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain.* Cornell Aeronautical Laboratory, Psychological Review, 1958. **65**(6): p. 386-408.
201. Cybenko, G.V., *Approximation by Superpositions of a Sigmoidal function.* Mathematics of Control, Signals and Systems, 1989. **2**(4): p. 303-314.
202. Vapnik, V. and A. Lerner, *Pattern Recognition using the Generalized Portrait Method.* Automation and Remote Control, 1963. **24**: p. 774-780.
203. Vapnik, V.N., *The nature of statistical learning theory*. 2000: springer.
204. Aizerman, M., E. Braverman, and L. Rozonoer, *Theoretical foundations of the potential function method in pattern recognition learning.* Automation and Remote Control, 1964. **25**: p. 821-837.

205. *ATP-binding cassette sub-family B member 1 [Homo sapiens], Calcein AM assay*. 2006, National Center for Biotechnology Information, PubChem.
206. Hollo, Z., et al., *Parallel functional and immunological detection of human multidrug resistance proteins, P-glycoprotein and MRP1.* Anticancer Res., 1998. **18**(4C): p. 2981-7.
207. Karaszi, E., et al., *Calcein assay for multidrug resistance reliably predicts therapy response and survival rate in acute myeloid leukaemia.* Br. J. Haematol., 2001. **112**(2): p. 308-14.
208. PDSP, *MDR-1*. 2006, NIMH Psychoactive Drug Screening Program.
209. Steinbeck, C., et al., *The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics.* J. Chem. Inf. Comput. Sci., 2003. **43**(2): p. 493-500.
210. Qian, X.D. and W.T. Beck, *Binding of an optically pure photoaffinity analogue of verapamil, LU-49888, to P-glycoprotein from multidrug-resistant human leukemic cell lines.* Cancer Res, 1990. **50**(4): p. 1132-7.
211. Doran, A., et al., *The impact of P-glycoprotein on the disposition of drugs targeted for indications of the central nervous system: evaluation using the MDR1A/1B knockout mouse model.* Drug Metab. Dispos., 2005. **33**(1): p. 165-74.
212. de Cerqueira Lima, P., et al., *Combinatorial QSAR modeling of P-glycoprotein substrates.* J. Chem. Inf. Model., 2006. **46**(3): p. 1245-54.
213. Chang, C., et al., *Pharmacophore-based discovery of ligands for drug transporters.* Adv. Drug Deliv. Rev., 2006. **58**(12-13): p. 1431-50.
214. Polli, J.W., et al., *Rational use of in vitro P-glycoprotein assays in drug discovery.* J. Pharmacol. Exp. Ther., 2001. **299**(2): p. 620-8.
215. Wang, R.B., et al., *Structure-activity relationship: analyses of p-glycoprotein substrates and inhibitors.* J. Clin. Pharm. Ther., 2003. **28**(3): p. 203-28.
216. Didziapetris, R., et al., *Classification analysis of P-glycoprotein substrate specificity.* J. Drug Target., 2003. **11**(7): p. 391-406.
217. Stouch, T.R. and O. Gudmundsson, *Progress in understanding the structure-activity relationships of P-glycoprotein.* Adv. Drug Deliv. Rev., 2002. **54**(3): p. 315-28.
218. Seelig, A. and E. Landwojtowicz, *Structure-activity relationship of P-glycoprotein substrates and modifiers.* Eur. J. Pharm. Sci., 2000. **12**(1): p. 31-40.
219. Zhou, S.F. and X. Lai, *An update on clinical drug interactions with the herbal antidepressant St. John's wort.* Curr. Drug Metab., 2008. **9**(5): p. 394-409.
220. Carson, S.W., A.D. Ousmanou, and S.L. Hoyler, *Emerging significance of P-glycoprotein in understanding drug disposition and drug interactions in psychopharmacology.* Psychopharmacol Bull, 2002. **36**(1): p. 67-81.
221. Zastre, J.A., et al., *Up-regulation of P-glycoprotein by HIV protease inhibitors in a human brain microvessel endothelial cell line.* J. Neurosci. Res., 2008. **87**(4): p. 1023-1036.
222. Breiman, L., *Bagging Predictors.* Mach. Learn., 1996. **24**(2): p. 123-140.
223. Huang, J., et al., *Identifying P-glycoprotein substrates using a support vector machine optimized by a particle swarm.* J. Chem. Inf. Model., 2007. **47**(4): p. 1638-47.
224. Wang, Y.H., et al., *Classification of substrates and inhibitors of P-glycoprotein using unsupervised machine learning approach.* J. Chem. Inf. Model., 2005. **45**(3): p. 750-7.
225. Wishart, D.S., et al., *DrugBank: a comprehensive resource for in silico drug discovery and exploration.* Nucleic Acids Res, 2006. **34**(Database issue): p. D668-72.
226. Vasanthanathan, P., et al., *Classification of cytochrome P450 1A2 inhibitors and noninhibitors by machine learning techniques.* Drug Metab Dispos, 2009. **37**(3): p. 658-64.
227. Yap, C.W. and Y.Z. Chen, *Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines.* J Chem Inf Model, 2005. **45**(4): p. 982-92.
228. Katritzky, A.R., et al., *Correlation of Boiling Points with Molecular Structure. 1. A Training Set of 298 Diverse Organics and a Test Set of 9 Simple Inorganics.* Journal of Physical Chemistry, 1996. **100**(24): p. 10400-10407.
229. Pearlman, R.S. and K.M. Smith, *Metric Validation and the Receptor-Relevant Subspace Concept.* J Chem Inf Comput Sci, 1999. **39**(1): p. 28-35.
230. Smith, D.A., M.J. Ackland, and B.C. Jones, *Properties of cytochrome P450 isoenzymes and their substrates part 2: properties of cytochrome P450 substrates.* Drug Discov Today, 1997. **2**(11): p. 479-486.
231. Lewis, D.F.V., *Structural characteristics of human P450s involved in drug metabolism: QSARs and lipophilicity profiles.* Toxicology, 2000. **144**(1-3): p. 197-203.
232. Langowski, J. and A. Long, *Computer systems for the prediction of xenobiotic metabolism.* Adv Drug Deliv Rev, 2002. **54**(3): p. 407-15.

233. Smith, D.A., M.J. Ackland, and B.C. Jones, *Properties of cytochrome P450 isoenzymes and their substrates part 1: active site characteristics.* Drug Discov Today, 1997. **2**(10): p. 406-414.
234. Mao, B., et al., *QSAR modeling of in vitro inhibition of cytochrome P450 3A4.* J Chem Inf Model, 2006. **46**(5): p. 2125-34.
235. Zhou, S.F., *Drugs behave as substrates, inhibitors and inducers of human cytochrome P450 3A4.* Curr Drug Metab, 2008. **9**(4): p. 310-22.
236. Stjernschantz, E., N.P. Vermeulen, and C. Oostenbrink, *Computational prediction of drug binding and rationalisation of selectivity towards cytochromes P450.* Expert Opin Drug Metab Toxicol, 2008. **4**(5): p. 513-27.
237. Yap, C.W., et al., *Application of support vector machines to in silico prediction of cytochrome p450 enzyme substrates and inhibitors.* Curr Top Med Chem, 2006. **6**(15): p. 1593-607.
238. Leong, M.K., et al., *Development of a new predictive model for interactions with human cytochrome P450 2A6 using pharmacophore ensemble/support vector machine (PhE/SVM) approach.* Pharm Res, 2009. **26**(4): p. 987-1000.
239. Leong, M.K. and T.H. Chen, *Prediction of cytochrome P450 2B6-substrate interactions using pharmacophore ensemble/support vector machine (PhE/SVM) approach.* Med Chem, 2008. **4**(4): p. 396-406.
240. Hammann, F., et al., *Development of decision tree models for substrates, inhibitors, and inducers of p-glycoprotein.* Curr Drug Metab, 2009. **10**(4): p. 339-46.
241. Andrade, R.J., et al., *Assessment of drug-induced hepatotoxicity in clinical practice: a challenge for gastroenterologists.* World J Gastroenterol, 2007. **13**(3): p. 329-40.
242. Ramachandran, R. and S. Kakar, *Histological patterns in drug-induced liver disease.* J Clin Pathol, 2009. **62**: p. 481-492.
243. Hammann, F., et al., *Development of Decision Tree Models for Substrates, Inhibitors, and Inducers of P-Glycoprotein.* Current Drug Metabolism, 2009. **10**(4).
244. Tang, W., *Drug metabolite profiling and elucidation of drug-induced hepatotoxicity.* Expert Opin Drug Metab Toxicol, 2007. **3**(3): p. 407-20.
245. Fukudo, M., et al., *Pharmacodynamic analysis of tacrolimus and cyclosporine in living-donor liver transplant patients.* Clin Pharmacol Ther, 2005. **78**(2): p. 168-81.
246. Swidsinski, A., et al., *Mucosal flora in inflammatory bowel disease.* Gastroenterology, 2002. **122**(1): p. 44-54.
247. Ardizzone, S., et al., *Extraintestinal manifestations of inflammatory bowel disease.* Dig Liver Dis, 2008. **40 Suppl 2**: p. S253-9.
248. Xavier, R.J. and D.K. Podolsky, *Unravelling the pathogenesis of inflammatory bowel disease.* Nature, 2007. **448**(7152): p. 427-34.
249. Torok, H.P., B. Goke, and A. Konrad, *Pharmacogenetics of Crohn's disease.* Pharmacogenomics, 2008. **9**(7): p. 881-93.
250. Huls, M., F.G. Russel, and R. Masereeuw, *The role of ATP binding cassette transporters in tissue defense and organ regeneration.* J Pharmacol Exp Ther, 2009. **328**(1): p. 3-9.
251. Cascorbi, I., *Role of pharmacogenetics of ATP-binding cassette transporters in the pharmacokinetics of drugs.* Pharmacol Ther, 2006. **112**(2): p. 457-73.
252. Zhao, R., et al., *Breast Cancer Resistance Protein Interacts with Various Compounds in vitro, but Plays a Minor Role in Substrate Efflux at the Blood-Brain Barrier.* Drug Metab Dispos, 2009.
253. Tai, L.M., et al., *P-glycoprotein and breast cancer resistance protein restrict apical-to-basolateral permeability of human brain endothelium to amyloid-beta.* J Cereb Blood Flow Metab, 2009.
254. Urcelay, E., et al., *MDR1 gene: susceptibility in Spanish Crohn's disease and ulcerative colitis patients.* Inflamm Bowel Dis, 2006. **12**(1): p. 33-7.
255. Fiedler, T., et al., *Possible role of MDR1 two-locus genotypes for young-age onset ulcerative colitis but not Crohn's disease.* Eur J Clin Pharmacol, 2007. **63**(10): p. 917-25.
256. Ho, G.T., et al., *Allelic variations of the multidrug resistance gene determine susceptibility and disease behavior in ulcerative colitis.* Gastroenterology, 2005. **128**(2): p. 288-96.
257. Ostergaard, M., et al., *Cyclooxygenase-2, multidrug resistance 1, and breast cancer resistance protein gene polymorphisms and inflammatory bowel disease in the Danish population.* Scand J Gastroenterol, 2009. **44**(1): p. 65-73.
258. Nikolaus, S. and S. Schreiber, *Diagnostics of inflammatory bowel disease.* Gastroenterology, 2007. **133**(5): p. 1670-89.
259. Korenaga, Y., et al., *Association of the BCRP C421A polymorphism with nonpapillary renal cell carcinoma.* Int J Cancer, 2005. **117**(3): p. 431-4.

260. Eap, C.B., et al., *CYP3A activity measured by the midazolam test is not related to 3435 C >T polymorphism in the multiple drug resistance transporter gene.* Pharmacogenetics, 2004. **14**(4): p. 255-60.
261. Palmieri, O., et al., *Multidrug resistance 1 gene polymorphisms are not associated with inflammatory bowel disease and response to therapy in Italian patients.* Aliment Pharmacol Ther, 2005. **22**(11-12): p. 1129-38.
262. Glas, J., et al., *MDR1 gene polymorphism in ulcerative colitis.* Gastroenterology, 2004. **126**(1): p. 367.
263. Potocnik, U., et al., *Polymorphisms in multidrug resistance 1 (MDR1) gene are associated with refractory Crohn disease and ulcerative colitis.* Genes Immun, 2004. **5**(7): p. 530-9.
264. Roses, A.D., *Pharmacogenetics in drug discovery and development: a translational perspective.* Nat Rev Drug Discov, 2008. **7**(10): p. 807-17.
265. Meech, R. and P.I. Mackenzie, *Structure and function of uridine diphosphate glucuronosyltransferases.* Clin Exp Pharmacol Physiol, 1997. **24**(12): p. 907-15.
266. Mackenzie, P.I., et al., *The UDP glycosyltransferase gene superfamily: recommended nomenclature update based on evolutionary divergence.* Pharmacogenetics, 1997. **7**(4): p. 255-69.
267. King, C.D., et al., *UDP-glucuronosyltransferases.* Curr Drug Metab, 2000. **1**(2): p. 143-61.
268. Crigler, J.F., Jr. and V.A. Najjar, *Congenital familial nonhemolytic jaundice with kernicterus.* Pediatrics, 1952. **10**(2): p. 169-80.
269. Arias, I.M., *Chronic unconjugated hyperbilirubinemia without overt signs of hemolysis in adolescents and adults.* J Clin Invest, 1962. **41**: p. 2233-45.
270. Bosma, P.J., et al., *Bilirubin UDP-glucuronosyltransferase 1 is the only relevant bilirubin glucuronidating isoform in man.* J Biol Chem, 1994. **269**(27): p. 17960-4.
271. Seppen, J., et al., *Discrimination between Crigler-Najjar type I and II by expression of mutant bilirubin uridine diphosphate-glucuronosyltransferase.* J Clin Invest, 1994. **94**(6): p. 2385-91.
272. Labrune, P., et al., *Crigler-Najjar type II disease inheritance: a family study.* J Inherit Metab Dis, 1989. **12**(3): p. 302-6.
273. Arias, I.M., et al., *Chronic nonhemolytic unconjugated hyperbilirubinemia with glucuronyl transferase deficiency. Clinical, biochemical, pharmacologic and genetic evidence for heterogeneity.* Am J Med, 1969. **47**(3): p. 395-409.
274. Ritter, J.K., et al., *Expression and inducibility of the human bilirubin UDP-glucuronosyltransferase UGT1A1 in liver and cultured primary hepatocytes: evidence for both genetic and environmental influences.* Hepatology, 1999. **30**(2): p. 476-84.
275. Pett, S. and A.P. Mowat, *Crigler-Najjar syndrome types I and II. Clinical experience--King's College Hospital 1972-1978. Phenobarbitone, phototherapy and liver transplantation.* Mol Aspects Med, 1987. **9**(5): p. 473-82.
276. Yaffe, S.J., et al., *Enhancement of glucuronide-conjugating capacity in a hyperbilirubinemic infant due to apparent enzyme induction by phenobarbital.* N Engl J Med, 1966. **275**(26): p. 1461-6.
277. Crigler, J.F., Jr. and N.I. Gold, *Effect of sodium phenobarbital on bilirubin metabolism in an infant with congenital, nonhemolytic, unconjugated hyperbilirubinemia, and kernicterus.* J Clin Invest, 1969. **48**(1): p. 42-55.
278. Ertel, I.J. and W.A. Newton, Jr., *Therapy in congenital hyperbilirubinemia: phenobarbital and diethylnicotinamide.* Pediatrics, 1969. **44**(1): p. 43-8.
279. Olinga, P., et al., *Coordinated induction of drug transporters and phase I and II metabolism in human liver slices.* Eur J Pharm Sci, 2008. **33**(4-5): p. 380-9.
280. Jemnitz, K., et al., *Glucuronidation of thyroxine in primary monolayer cultures of rat hepatocytes: in vitro induction of UDP-glucuronosyltranferases by methylcholanthrene, clofibrate, and dexamethasone alone and in combination.* Drug Metab Dispos, 2000. **28**(1): p. 34-7.
281. Jeong, H., et al., *Regulation of UDP-glucuronosyltransferase (UGT) 1A1 by progesterone and its impact on labetalol elimination.* Xenobiotica, 2008. **38**(1): p. 62-75.
282. Martin, P., et al., *Comparison of the induction profile for drug disposition proteins by typical nuclear receptor activators in human hepatic and intestinal cells.* Br J Pharmacol, 2008. **153**(4): p. 805-19.
283. Kwan, P. and M.J. Brodie, *Neuropsychological effects of epilepsy and antiepileptic drugs.* Lancet, 2001. **357**(9251): p. 216-22.
284. Gillham, R.A., et al., *Cognitive function in adult epileptic patients established on anticonvulsant monotherapy.* Epilepsy Res, 1990. **7**(3): p. 219-25.
285. Linde, K., et al., *St John's wort for depression--an overview and meta-analysis of randomised clinical trials.* Bmj, 1996. **313**(7052): p. 253-8.

286. Moore, L.B., et al., *St. John's wort induces hepatic drug metabolism through activation of the pregnane X receptor.* Proc Natl Acad Sci U S A, 2000. **97**(13): p. 7500-2.
287. Godtel-Armbrust, U., et al., *Variability in PXR-mediated induction of CYP3A4 by commercial preparations and dry extracts of St. John's wort.* Naunyn Schmiedebergs Arch Pharmacol, 2007. **375**(6): p. 377-82.
288. Mueller, S.C., et al., *The extent of induction of CYP3A by St. John's wort varies among products and is linked to hyperforin dose.* Eur J Clin Pharmacol, 2006. **62**(1): p. 29-36.
289. Kadakol, A., et al., *Genetic lesions of bilirubin uridine-diphosphoglucuronate glucuronosyltransferase (UGT1A1) causing Crigler-Najjar and Gilbert syndromes: correlation of genotype to phenotype.* Hum Mutat, 2000. **16**(4): p. 297-306.
290. Aono, S., et al., *Identification of defect in the genes for bilirubin UDP-glucuronosyl-transferase in a patient with Crigler-Najjar syndrome type II.* Biochem Biophys Res Commun, 1993. **197**(3): p. 1239-44.
291. Wang, Z., et al., *The effects of St John's wort (Hypericum perforatum) on human cytochrome P450 activity.* Clin Pharmacol Ther, 2001. **70**(4): p. 317-26.
292. Gurley, B.J., et al., *Cytochrome P450 phenotypic ratios for predicting herb-drug interactions in humans.* Clin Pharmacol Ther, 2002. **72**(3): p. 276-87.
293. Dresser, G.K., et al., *Coordinate induction of both cytochrome P4503A and MDR1 by St John's wort in healthy subjects.* Clin Pharmacol Ther, 2003. **73**(1): p. 41-50.
294. Link, B., et al., *Determination of midazolam and its hydroxy metabolites in human plasma and oral fluid by liquid chromatography/electrospray ionization ion trap tandem mass spectrometry.* Rapid Commun Mass Spectrom, 2007. **21**(9): p. 1531-40.
295. Zhu, B., et al., *Characterization of 1'-hydroxymidazolam glucuronidation in human liver microsomes.* Drug Metab Dispos, 2008. **36**(2): p. 331-8.
296. Klieber, S., et al., *Contribution of the N-glucuronidation pathway to the overall in vitro metabolic clearance of midazolam in humans.* Drug Metab Dispos, 2008. **36**(5): p. 851-62.
297. Erkul, I., H. Yavuz, and A. Ozel, *Clofibrate treatment of neonatal jaundice.* Pediatrics, 1991. **88**(6): p. 1292-4.
298. *W.H.O. cooperative trial on primary prevention of ischaemic heart disease using clofibrate to lower serum cholesterol: mortality follow-up. Report of the Committee of Principal Investigators.* Lancet, 1980. **2**(8191): p. 379-85.
299. Ernst, E., et al., *Adverse effects profile of the herbal antidepressant St. John's wort (Hypericum perforatum L.).* Eur J Clin Pharmacol, 1998. **54**(8): p. 589-94.
300. Park, J., et al., *The role of oxygen in the antiviral activity of hypericin and hypocrellin.* Photochem Photobiol, 1998. **68**(4): p. 593-7.
301. Kubin, A., et al., *Hypericin - the facts about a controversial agent.* Curr Pharm Des, 2005. **11**(2): p. 233-53.
302. Schey, K.L., et al., *Photooxidation of lens alpha-crystallin by hypericin (active ingredient in St. John's Wort).* Photochem Photobiol, 2000. **72**(2): p. 200-3.
303. Schempp, C.M., et al., *Effect of oral administration of Hypericum perforatum extract (St. John's Wort) on skin erythema and pigmentation induced by UVB, UVA, visible light and solar simulated radiation.* Phytother Res, 2003. **17**(2): p. 141-6.
304. Schmitt, L.A., et al., *Reduction in hypericin-induced phototoxicity by Hypericum perforatum extracts and pure compounds.* J Photochem Photobiol B, 2006. **85**(2): p. 118-30.
305. Agusti, A. and J.B. Soriano, *COPD as a systemic disease.* Copd, 2008. **5**(2): p. 133-8.
306. Arcavi, L. and N.L. Benowitz, *Cigarette smoking and infection.* Arch Intern Med, 2004. **164**(20): p. 2206-16.
307. Gerszten, R.E. and T.J. Wang, *The search for new cardiovascular biomarkers.* Nature, 2008. **451**(7181): p. 949-52.
308. Haussmann, H.J., *Smoking and lung cancer: future research directions.* Int J Toxicol, 2007. **26**(4): p. 353-64.
309. Vellappally, S., et al., *Smoking related systemic and oral diseases.* Acta Medica (Hradec Kralove), 2007. **50**(3): p. 161-6.
310. Mackay, J., M. Ericksen, and O. Shafey, *The tobacco atlas*. 2nd ed. 2006, Atlanta: American Cancer Society.
311. Benowitz, N.L., *Clinical pharmacology of nicotine: implications for understanding, preventing, and treating tobacco addiction.* Clin Pharmacol Ther, 2008. **83**(4): p. 531-41.
312. Lillington, G.A., C.T. Leonard, and D.P. Sachs, *Smoking cessation. Techniques and benefits.* Clin Chest Med, 2000. **21**(1): p. 199-208, xi.

313. Henningfield, J.E. and R.M. Keenan, *Nicotine delivery kinetics and abuse liability.* J Consult Clin Psychol, 1993. **61**(5): p. 743-50.
314. Hatsukami, D.K., L.F. Stead, and P.C. Gupta, *Tobacco addiction.* Lancet, 2008. **371**(9629): p. 2027-38.
315. Jorenby, D.E., et al., *Efficacy of varenicline, an alpha4beta2 nicotinic acetylcholine receptor partial agonist, vs placebo or sustained-release bupropion for smoking cessation: a randomized controlled trial.* Jama, 2006. **296**(1): p. 56-63.
316. Eisenberg, M.J., et al., *Pharmacotherapies for smoking cessation: a meta-analysis of randomized controlled trials.* Cmaj, 2008. **179**(2): p. 135-44.
317. Tang, J.L., M. Law, and N. Wald, *How effective is nicotine replacement therapy in helping people to stop smoking?* Bmj, 1994. **308**(6920): p. 21-6.
318. Cheng, Y.H., et al., *Development of a novel nasal nicotine formulation comprising an optimal pulsatile and sustained plasma nicotine profile for smoking cessation.* J Control Release, 2002. **79**(1-3): p. 243-54.
319. Bussemer, T., I. Otto, and R. Bodmeier, *Pulsatile drug-delivery systems.* Crit Rev Ther Drug Carrier Syst, 2001. **18**(5): p. 433-58.
320. Lemmer, B., *Circadian rhythms and drug delivery.* J Control Release, 1991. **16**(1-2): p. 63-74.
321. Isaac, P.F. and M.J. Rand, *Cigarette smoking and plasma levels of nicotine.* Nature, 1972. **236**(5345): p. 308-10.
322. Smith, T.A., et al., *Nicotine patch therapy in adolescent smokers.* Pediatrics, 1996. **98**(4 Pt 1): p. 659-67.
323. Flood, P. and D. Daniel, *Intranasal nicotine for postoperative pain treatment.* Anesthesiology, 2004. **101**(6): p. 1417-21.
324. Hong, D., et al., *Transdermal Nicotine Patch for Postoperative Pain Management: A Pilot Dose-Ranging Study.* Anesthesia and Analgesia, 2008. **107**(3): p. 1005-1010.
325. Ocak, F. and I. Agabeyoglu, *Development of a membrane-controlled transdermal therapeutic system containing isosorbide dinitrate.* Int J Pharm, 1999. **180**(2): p. 177-83.
326. Maurer, H.H., *Position of chromatographic techniques in screening for detection of drugs or poisons in clinical and forensic toxicology and/or doping control.* Clin Chem Lab Med, 2004. **42**(11): p. 1310-1324.
327. Pfleger, K., et al., *Mass spectral and GC data of drugs, poisons, pesticides, pollutants, and their metabolites*. 1992: VCH New York.
328. Theobald, D.S. and H.H. Maurer, *Studies on the metabolism and toxicological detection of the designer drug 2,5-dimethoxy-4-methyl-beta- phenethylamine (2C-D) in rat urine using gas chromatographic/mass spectrometric techniques.* J Mass Spectrom, 2006. **41**(11): p. 1509-1519.
329. Dussy, F.E., C. Hamberg, and T.A. Briellmann, *Quantification of benzodiazepines in whole blood and serum.* Int J Leg Med, 2006. **120**(6): p. 323-330.
330. Lambert, W.E., J.F. Van Bocxlaer, and A.P. De Leenheer, *Potential of high-performance liquid chromatography with photodiode array detection in forensic toxicology.* J Chromatogr B Biomed Sci Appl, 1997. **689**(1): p. 45-53.
331. Maier, R.D. and M. Bogusz, *Identification power of a standardized HPLC-DAD system for systematic toxicological analysis.* J Anal Toxicol, 1995. **19**(2): p. 79-83.
332. Tracqui, A., P. Kintz, and P. Mangin, *Systematic Toxicological Analysis using HPLC/DAD.* J. Forensic Sci., 1995. **40**(2): p. 254-262.
333. Decaestecker, T.N., et al., *Evaluation of automated single mass spectrometry to tandem mass spectrometry function switching for comprehensive drug profiling analysis using a quadrupole time-of-flight mass spectrometer.* Rapid Commun Mass Spectrom, 2000. **14**(19): p. 1787-1792.
334. Decaestecker, T.N., et al., *Information-dependent acquisition-mediated LC-MS/MS screening procedure with semiquantitative potential.* Anal Chem, 2004. **76**(21): p. 6365-6373.
335. Fitzgerald, R.L., J.D. Rivera, and D.A. Herold, *Broad spectrum drug identification directly from urine, using liquid chromatography-tandem mass spectrometry.* Clin Chem, 1999. **45**(8 Pt 1): p. 1224-1234.
336. Gergov, M., I. Ojanpera, and E. Vuori, *Simultaneous screening for 238 drugs in blood by liquid chromatography-ion spray tandem mass spectrometry with multiple-reaction monitoring.* J Chromatogr B Analyt Technol Biomed Life Sci, 2003. **795**(1): p. 41-53.
337. Marquet, P., N. Venisse, and E. Lacassie, Analysis, 2000. **28**(41).
338. Maurer, H.H., *Analytical toxicology.* Anal Bioanal Chem, 2007. **388**(7): p. 1311-1311.
339. Venisse, N., et al., *A general unknown screening procedure for drugs and toxic compounds in serum using liquid chromatography-electrospray-single quadrupole mass spectrometry.* J Anal Toxicol, 2003. **27**(1): p. 7-14.

340. Weinmann, W., et al., *Screening for drugs in serum by electrospray ionization/collision-induced dissociation and library searching.* J Am Soc Mass Spectrom, 1999. **10**(10): p. 1028-1037.
341. Maurer, H.H., et al., *Screening for library-assisted identification and fully validated quantification of 22 beta-blockers in blood plasma by liquid chromatography-mass spectrometry with atmospheric pressure chemical ionization.* J Chromatogr A, 2004. **1058**(1-2): p. 169-181.
342. Sauvage, F.-L., et al., *Screening of drugs and toxic compounds with liquid chromatography-linear ion trap tandem mass spectrometry.* Clin Chem, 2006. **52**(9): p. 1735-1742.
343. Lee, M.S. and E.H. Kerns, *LC/MS applications in drug development.* Mass Spectrom Rev, 1999. **18**(3-4): p. 187-279.
344. Maurer, H.H., *In: Bogusz M (ed) Handbook of analytical separation sciences: forensic sciences*. 2007, Elsevier, Amsterdam.
345. Dams, R., et al., *Matrix effect in bio-analysis of illicit drugs with LC-MS/MS: influence of ionization type, sample preparation, and biofluid.* J Am Soc Mass Spectrom, 2003. **14**(11): p. 1290-1294.
346. Souverain, S., S. Rudaz, and J.-L. Veuthey, *Matrix effect in LC-ESI-MS and LC-APCI-MS with off-line and on-line extraction procedures.* J Chromatogr A, 2004. **1058**(1-2): p. 61-66.
347. Chih-Chung, C. and L. Chih-Jen, *LIBSVM: a library for support vector machines*. 2001.

Curriculum Vitae

Personal information

Full name:	Felix Georg Michael Hammann
Date and place of birth:	8th of March 1978, Frankfurt am Main (Germany)
Nationality:	German
Home address:	Hebelstrasse 82 4056 Basel Telephone: +41 61 261 1606
Institutional address:	Department of Internal Medicine / Division of Clinical Pharmacology &Toxicology Department of Gastroenterology University Hospital Basel Hebelstrasse 2 4031 Basel Telephone: +41 61 328 7742 E-Mail: felix.hammann@unibas.ch

Education

2008:	Medical thesis “Pharmacodynamics of oral oxycodone in the presence of the CYP2D6 inhibitor paroxetine or the CYP3A4 inhibitor ketoconazole” Division of Clinical Pharmacology and Toxicology (University Hospital Basel, supervisor: Prof. Dr. Jürgen Drewe)
2007 – now:	Ph.D. student Molecular transporters Division of Clinical Pharmacology and Toxicology (University Hospital Basel, supervisor: Prof. Dr. Jürgen Drewe)
1998 – 2005:	Study of Medicine and Biology University of Basel, Switzerland Federal Swiss diploma in Human Medicine

Curriculum Vitae

1984 – 1997:	Primary and Secondary School Kelsterbach, Germany; Rheinfelden, Switzerland; Ossining, NY, USA; Muttenz, Switzerland

Grants

2007:	Altana-Fonds of the Departement of Internal Medicine (University Hospital Basel) Determination of a genetic mutation in patients with inflammatory bowel disease

Publications

Hammann F., Gutmann H., Baumann U., Helma C., Drewe J., *Classification of Cytochrome P450 Activities Using Machine Learning Methods,* Mol. Pharm. 2009 Nov-Dec;6(6):1920-6.

Hammann F., Gutmann H., Jecklin U., Maunz A., Helma C., Drewe J., *Development of Decision Tree Models for Substrates, Inhibitors, and Inducers of P-Glycoprotein*, Curr. Drug Metab. 2009 May;10(4):339-46.

Kummer O., Haschke M., **Hammann F.**, Bodmer M., Bruderer S., Regnault Y., Dingemanse J., Krähenbühl S., *Comparison of the dissolution and pharmacokinetic profiles of two galenical formulations of the endothelin receptor antagonist macitentan*, Eur. J. Pharm. Sci. 2009 Nov 5;38(4):384-8.

Kummer O., **Hammann F.**, Krähenbühl S., Bodmer M., *Medikamentös-toxische Hepatitis*, Schweiz. Rundsch. Med. Prax., 2008;97:235-241

Gutmann H., Hruz P., Zimmermann C., Straumann A., Terracciano L., **Hammann F.**, Lehmann F., Beglinger C., Drewe J., *Breast cancer resistance protein (BCRP) and P-glycoprotein (P-gp) expression in newly diagnosed and therapy refractory ulcerative colitis*, Digestion, 2008;3;78(2-3):154-162

Kummer O., Novakova K., Burkhard T., **Hammann F.**, Krähenbühl S., Bodmer M., *Medikamenten-induzierte mikroskopische Kolitis*, Schweiz. Rundsch. Med. Prax., 2007;96:1293-1297

DoseAdapt, eine Webanwendung zur Dosisanpassung bei Niereninsuffizienz (Extension and maintenance)

Poster Presentations

Effect of CYP3A4 or CYP2D6 inhibition on the kinetics and dynamics of oxycodone, 77th annual meeting of the Swiss Society for Internal Medicine, Basel, Switzerland, May 2009

Successful treatment of a patient with Crigler Najjar type II syndrome with St. John's wort, 77th annual meeting of the Swiss Society for Internal Medicine, Basel, Switzerland, May 2009

Attended Meetings

10. Blut-Hirnschranken-Expertentreffen, May, 2008, Bad Herrenalb, Germany

During my studies, I attended lectures by:

Arber S., Beier K., Betz G., Bickle T.A., Bienz K.A., Boelsterli U., Boller Th., Drewe J., Eberle A., Engel J., Erb P., Ernst B., Folkers G., Gehring J., Gescheidt G., Grzesiek S., Guentert T., Güntherodt H.-J., Hauri H.-P., Hauser P., Hersberger K., Huwyler J., Im Hof H.-C., Imanidis G., Jenal U., Keller W., Kessler M., Kiefhaber T., Körner C., Krähenbühl S., Kunz D., Leuenberger H., Lüdin E., Lüthi A., Mayans O., Meier B., Meier C., Meier T., Melchers F., Meyer J., Meyer U., Monard D., Mühlebach S., Müller H., Müller J., Neri D., Otten U., Pfleiderer G., Rehmann-Sutter C., Reichert H., Rüegg M., Schaffner W., Schirmer T., Schlienger R., Schmid B., Schmid V., Scholer A., Schönenberger C., Seelig J., Séquin U., Sick I., Spiess M., Spornitz U., Stoeckli E., Strazewski P., Vedani A., Vetter T., Wessels H.-P., Zaugg C., Zuberbühler A.