In Silico Prediction of Drug Transport Across Physiological Barriers

Inauguraldissertation

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

von

Claudia Suenderhauf
aus Untereggen, St.Gallen

Basel, 2011

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät

auf Antrag von

Prof. Dr. Jörg Huwyler

Prof. Dr. Jürgen Drewe

Basel, den 21. Juni 2011

Prof. Dr. Martin Spiess

Dekan

Für Richy

Acknowledgments

Als erstes möchte ich mich bei meinem Doktorvater Prof. Dr. Jörg Huwyler bedanken, der es mir ermöglichte in seiner Gruppe ein PhD Studium zu absolvieren. Seine enthusiastische und konstruktive Art half mir stets scheinbar unlösbare Probleme zu bewältigen und wieder mit neuem Mut an die Arbeit zu gehen. Unter seiner Leitung erwarb ich mir eine grosse Selbständigkeit im wissenschaftlichen Arbeiten. Für diese einzigartigen Erfahrungen und Jahre bei Ihm möchte ich mich herzlich bedanken.

Mein tiefster Dank gebührt Dr. Felix Hammann, der mich als Supervisor und Freund durch das Auf und Ab der letzten 3 Jahre meiner Doktorarbeit auf allen Ebenen unterstützte, betreute und förderte. Ich hätte mir keinen besseren Betreuer für mich vorstellen können.

Des Weiteren möchte ich mich bei Prof. Dr. Jürgen Drewe bedanken. Sein Rat und aussergewöhnliches Engagement trugen massgeblich dazu bei, dass ich meine Thesis zu einem guten Ende bringen konnte. Es war für mich eine Ehre, ihn in meinem PhD Komitee haben zu dürfen. Ich möchte mich auch ganz herzlich bei Prof. Dr. Angelo Vedani bedanken, der sich als Chairman für meine Defense zur Verfügung stellte. Ich hätte nicht gedacht, dass eine Prüfung ein so erfreuliches Erlebnis sein kann. Dafür möchte ich mich herzlich bei ihnen beiden bedanken.

Dr. Elizaveta Fasler-Kan möchte ich herzlich danken für ihre enorme Unterstützung auf dem Gebiet der Zellkultur und des Imagings, vorallem während meiner medizinischen Doktorarbeit. Nicht zuletzt aber auch für die vielen guten (nicht immer nur wissenschaftlichen) Gespräche und Kaffeepausen. Auch Fabienne Thönen möchte ich an dieser Stelle ganz herzlich danken. Die vielen Gespräche über Zellkultur und andere Probleme haben mich durch die nicht immer einfachen Zeiten getragen.

Ein ganz grosses Dankeschön geht an Dr. Heike Gutmann und Dr. Christoph Helma für ihre liebe Unterstützung, Korrekturlesen und Hilfe - und nicht zu letzt für das "rescue-beer". Auch möchte ich mich auch ganz herzlich bei Mark Bamford für das Korrekturlesen bedanken.

Ganz besonderer Dank gebührt meinem Partner Gianrichy, der mir mit einer unglaublichen Liebe und Geduld in allen Hoch und Tiefs der letzten Jahre beistand. Zutiefst dankbar bin ich meiner Familie, Urs, Maja, Marco und Peter, dass sie mich immer unterstützt und an mich geglaubt haben. Auch der Familie Giamboi möchte ich aus ganzem Herzen danken, da sie für mich eine zweite Familie hier in Basel sind. Auch möchte ich Lilith und Grishnakh nicht unerwähnt lassen, die mich immer wieder auf tierische Art und Weise aus meinem Elfenbeinturm holten.

Bei der Schweizerischen Akademie der Medizinischen Wissenschaften (SAMW) und dem Schweizerischen National Fonds (SNF) möchte ich mich bedanken für die Aufnahme in das nationale MD/PhD Programm und dessen Finanzierung.

Abbreviations

%ABS	Absorption ratio
ABC	ATP binding cassette
ACE	Angiotensin converting enzyme
ACO	Ant colony optimization
aLogP	Octanol Water partition coefficient as defined by Ghose and Crippen
ANN	Artificial neural network
ASCII	American standard code for information interchange
ATP	Adenosine triphosphate
AUC	Area under the curve
BBB	Blood brain barrier
BCRP	Breast cancer resistance protein
BCUTS	Highest eigenvalue weighted for lowest atomic weight in the Burden matrix
BFS	Best first feature selection
CART	Classification and regression tree
CATS2D 02 LL	Distance of lipophilic pharmacophore groups at lag 2
CC	Correlation coefficient
CCR	Corrected classification rate
CDK	Chemical Development Kit
CFS	Linear correlation feature set
CHAID	Chi squared automatic interaction detector
cLogP	Computed partition coefficient
CNS	Central nervous system
cPSA	Computed polar surface area
CV	Cross-validation
CYP450	Cytochrome P450
dPSA2	Partial positive surface area multiplied by total positive charge on the molecule
DTI	Decision tree induction
DTIS	Decision tree induction feature set
Eig07_AEA(bo)	Eigenvalue number 7 from the augmented edge adjacency matrix weighted by bond order
FDA	Food and drug administration
fPSA	Charge weighted partial positive surface area divided by total molecular surface area
GATS6m	Geary's 2D autocorrelation matrices at lag 7 weighted by molecular mass
GATS7i	Geary's 2D autocorrelation matrices at lag 7 weighted by their ionization potential
GIT	Gastrointestinal tract
HB	Hemoglobin
HIV	Human immunodeficiency virus
i.v.	Intravenous application
Kier 2	Kappa shape index 2
KNN	K-nearest neighbor
LogBB	Logarithm of blood/brain partition measurement
LogP	Octanol water partition coefficient
LogPS	Logarithm of *in vivo* blood brain barrier permeability-surface area product
LOO	Leave-one-out
MATS7e	Moran's 2D autocorrelation matrices at lag 7 weighted by electronegativity
MATS7i	Moran's 2D autocorrelation matrices at lag 7 weighted by their ionization potential

MCC	Matthews correlation coefficient
ML	Machine learning
Mor10s	3DMoRSE descriptor weighted by intrinsic state at lag 10
Mor27p	3DMoRSE descriptor weighted by polarizability at lag 27
Mor28s	MoRSE3D descriptors at lag 28
MP	Multilayer perceptron
MRP	Multi drug resistance protein
MRP2	Multidrug resistance protein 2
NP	Nondeterministic polynomial
OATP	Organic anion transporting polypeptides
P_VSA_p2	VSA-like descriptor weighted for polarizability at lag 2
P-gp	P-Glycoprotein
Pf	*Plasmodium falciparum*
PK	Pharmacokinetics
pPSA2	Difference of pPSA2 divided by molecular surface and partial negative surface area multiplied
PS	Permeability surface product
PSA	Polar surface area
QSAR	Quantitative structure-activity relationship
QSPR	Quantitative structure-property relationship
R2	Coefficient of determination
R4s	R autocorrelation at lag 4 weighted by intrinsic state from the GETAWAY descriptors
Rbf	Radial basis function
RF	Random Forest
RMSE	Root mean squared error
ROC	Receiver operating characteristics
rPCG	Relative positive charge
SMARTS	SMILES arbitrary target identification
SMILES	Simplified molecular input line entry system
SpDiam_AEA(dm)	Spectral diameter from augmented edge adjacency (AEA) matrix, weighted by dipole moment
SpMAD EA [bo]	Spectral mean absolute deviation from the edge adjacency matrix, weighted by bond order
SVM	Support vector machines
TDB05p	Three-dimensional autocorrelation weighted for polarizability
tPSA	Topological polar surface area as defined by Ertl
UGT	Uridine 5'-diphospho-glucuronosyltransferases
xLogP	Octanol water partition coefficient as defined by Moriguchi

Table of contents

1 Summary of thesis

Physiological barriers maintain and safeguard homeostasis of certain body compartments by an increased resistance against free diffusion. Distribution and pharmacokinetics of drugs can be altered as well, if they have to cross these barriers in order to reach their target. Knowledge of the physicochemical and structural requirements for drug permeation is a key topic in drug design, development, and clinical application.

To assess processes on cellular barriers, *in vitro* methods are usually applied to elucidate single transport mechanisms or to study isolated transport. As the pharmacokinetics of a living system are often more complex and composed by a concatenation of several barriers, *in vivo* methods are required. However, this time consuming and expensive testing is not suited to answer the need for high-throughput screening of thousands of compounds in chemical databases. For these purpose *in silico* methods are ideally suited, which produce computational models to predict pharmacokinetics, drug distribution, or transport across single barriers. As these models are information compressions, they can give by themselves new insights into the process they predict.

In the present thesis, *in silico* models were developed to predict intestinal absorption, blood brain barrier permeation, drug permeation into breast milk, and active drug transport by the ATP binding cassette (ABC) transporter MRP2. In addition, a nature inspired modeling paradigm, ant colony optimization, was adapted and applied in the field of antimalaria drug therapy. These projects can be summarized as follows:

The first project concerned the modeling of human intestinal absorption. After oral administration and intestinal dissolution, a drug has to cross the gut wall in order to become available for the body. The process is mostly determined by passive diffusion and active transport. Active export and import of molecules on the enterocyte is regulated by a multitude of transport proteins and metabolic enzymes. A dataset of small drug-like compounds, on which information on their human intestinal absorption was available, was collected. Models trained on these data predicted human intestinal absorption with high accuracy. Several machine learning methods were compared as well as different feature sets. The features used to predict intestinal absorption resembled those known from modeling passive diffusion, which are measures of charge and lipophilicity. The models revealed also less commonly used descriptors to model human intestinal absorption, such as gravitational indices and moments of inertia.

The aim of the second project was to develop computational models to predict blood brain barrier (BBB) permeation. Development of new central nervous system (CNS) active drugs is hampered by limited brain permeation. As invasive methods have proven themselves to be ineffective and risky for patients, systemic application is the preferred route for drug administration into the brain. Hence, BBB permeability is a feature absolutely mandatory for any drug, which targets the CNS. Limited passive diffusion and active efflux and influx systems account for the complexity of this highly regulated barrier. To establish our models, a database of 163 compounds with information on the *in vivo* surface permeability product (LogPS) in rats was collected. Decision trees performed with high accuracy (CCR of 90.9 - 93.9%.) and revealed descriptors of lipophilicity and charge, which were yet described in models of passive BBB permeation. However, other descriptors as measures for molecular geometry and connectivity could be related to an active drug transport component. Moreover, a fragment-based approach indicated the involvement of stereochemistry to predict LogPS values.

The third project explores the physicochemical and structural requirements for drugs to pass from maternal blood into breast milk. While experimental assessment in humans is limited, computational methods are appropriate to model drug permeation into breast milk. Data preparation for these models was a challenging endeavor. Endpoints were reported in imprecise ways, which asked for a careful selection and binning of the instances. Despite these facts, the 10-fold cross-validated decision trees predicted the endpoint with high accuracy (CCR: 85.3 - 95.3%). Prominent descriptors were measures of molecular size, branching, charge and geometry. Importance of polar fragments was revealed by a fragment-based analysis.

The efflux transporter MRP2, a member of the ABC transporter family, was subject of the fourth study. Efflux transporters contribute substantially to barrier function by extruding potentially toxic substances. Three datasets were assembled from literature for MRP2 substrates, inducers, and inhibitors. For inducers and inhibitors, decision trees with high accuracy were grown. However, the substrate dataset did not qualify for decision tree induction, due to an underrepresentation of negative instances.

The fifth project deals with an ant colony optimization (ACO) algorithm, which was adapted for fragment based feature selection. The paradigm was tested to predict antimalarial activity of molecules. ACO was able to reveal chemical substructures characterizing antimalarial drug activity, which comprised passive diffusion through the erythrocyte membrane and parasite toxicity. The paradigm outperformed other algorithms such as decision trees or artificial neural networks on the same dataset.

2 Aim of thesis

Drugs have to cross several physiological barriers in the body in order to reach their target. Some of these barriers consist of specialized cells, which can exhibit increased tight connections between each other to reduce free diffusion. At these cell layers molecules can be actively transported with and against concentration gradients by a multitude of transport proteins. Barriers are found in the intestinal wall, the central nervous system, and the lactating breast epithelium. While they help to maintain homeostasis within the body and prevent permeation of toxic substances, these barriers can also substantially alter drug distribution or even completely prevent access to the site of action. It was therefore the aim of the present work to develop computational models using modern machine learning methods to predict drug permeation across physiological barriers.

We initially assessed human intestinal absorption using computational methods. After oral administration and intestinal dissolution, a drug has to cross the gut wall in order to become available for the body. Knowledge of intestinal absorption capacity is desirable as low intestinal absorption of a drug may limit its clinical application.

The second project aims to create methods to predict drug brain penetration, which is substantially restricted by the blood brain barrier. Knowledge on blood brain barrier permeation is therefore critical to develop drugs, which target the central nervous system.

The aim of the third project was to explore physicochemical and structural requirements for drug passage from maternal blood into breast milk. This topic is of particular relevance for drug safety in nursing. As ethical constraints limit in vivo experiments, computational methods are ideally suited to model this endpoint.

It was the aim of the fourth project to study a representative of the ABC transporter family, MRP2, as efflux transporters contribute substantially in maintaining barrier functions.

In the final study, we aimed to adapt an ant colony optimization algorithm to perform a fragment based feature selection. The paradigm was tested on the highly combined endpoint of antimalarial drug activity, which comprises passive diffusion through the erythrocyte membrane and toxic action on the parasite.

3 Introduction

3.1 A historical perspective

The use of drugs is as old as mankind. In fact, the use of herbal medicines might even predate modern *homo sapiens*. Findings of various different medicinal plants in Neanderthal tombs (60 000 years BC) indicate their use as remedies.[1] The 5300 year old “Oetzi” or “iceman” found in the Tyrolean Alps was carrying two pieces of birch fungus (*Pitoporus betulinus*) with him. It is nowadays believed that he knew of its beneficial effects (antibiotic and anti-inflammatory) and that it served him as an early first-aid kit.[2-4]

Despite its long history, drug discovery as a systematic, scientific, and multidisciplinary endeavor exists not much longer than a century. A dramatic development in chemistry induced a quantum leap of pharmaceutical sciences in the 19th century: The benzene theory formulated by Auguste Kekulé in 1865 led to intensive research on coal-tar derivatives, especially for their use as dyes.

The application of dyes inspired medical and pharmaceutical science. Paul Ehrlich discovered in the early 19th century a selective affinity of dyes for biological tissues. His observations led him to postulate the existence of “chemoreceptors” that should be exploited as therapeutic targets. With his statement “*Corpora non agunt nisi fixata*”, he was the first to formulate a basic principle of modern pharmacology. Namely, that active components have to bind their corresponding molecular target structure in order to cause a specific action. This theory was further refined by Emil Fischer (Key-lock principle, 1890) and Daniel E. Koshland (induced fit concept).[5] It became clear that a drug candidate should exhibit high target selectivity in order to be a good therapeutic. On the other hand, unspecific binding made a drug more prone to cause unwanted or toxic side effects.

Although, knowledge on target structures grew during the first decades of the 20th century, the greatest “block buster” drugs were still discovered by serendipitous accidents. The most famous example is probably the discovery of penicillin by Sir Alexander Fleming due to a fungus contamination of his bacterial cultures. His discovery conquered some of mankind's most ancient scourges, including syphilis, gangrene and tuberculosis. The more targeted identification of specific sites of action led also to remarkable results. William Campell for example isolated the avermectins from a soil sample collected from a golf course in Japan, which proved powerful against parasites. From systematic series of compounds, the semi-synthetic ivermectin turned out to be the most effective drug and was marketed ever since.[6] Another example of a success story was the development of Cyclosporine A. The immunosuppressive effect of the drug was discovered in a screening test developed by Hartmann F. Stähelin in Basel.[7]

In the late 1970s, genomic science led to a fast identification of drug target structures. *In vitro* assays were developed to quickly screen compounds for specific pharmacological properties. An automation of these experiments allowed for high-throughput screening, where thousands of compounds could be screened on one day. Despite the initial euphoria, its success stayed far behind expectations: although the number of molecules tested rose from 200 000 in 1990 to over 50 Million in 2000, the productivity of pharmaceutical industry, with respect of bringing new drugs to marked, could not be improved ever since.[8]

Many of these high-troughput screened compounds failed in the late and costly stage of drug development due to their unexpected or unfavorable pharmacokinetic behavior. The efficacy and safety of a drug is detrimentally dependent on its absorption characteristics, its tissue distribution, metabolism, as well as its excretion. A quick metabolism and elimination of a drug could abolish any therapeutic concentration on the target site, while a slow clearance leading to high plasma levels could cause toxic side effects. Several cellular and biochemical barriers can hamper distribution into body compartments and make predicting drug pharmacokinetics a challenging endeavor.

3.2 Pharmacokinetics in Drug Discovery

Depending on the application route and formulation, pharmacokinetics and bioavailability of a drug can substantially vary. The preferred route of administration is per oral since it is safe, cost-effective, and associated with high patients compliance. Low intestinal absorption of a drug may limit its clinical application, except in settings where the compounds target lies within the gastro-intestinal lumen (e.g., vancomycin, mesalazine). However, most orally applied drugs have to cross the intestinal epithelium and will be exposed to hepatic metabolism before reaching their site of action.

Limiting factors for intestinal drug absorption include low solubility or chemical instability in the gastrointestinal tract (GIT), high gastrointestinal metabolism, and poor intestinal membrane permeability.[9] Absorption kinetics are highly dependent on a compound’s solubility and hence galenic formulation, which influences exact location of dosage form disintegration in the GIT.[10] After intestinal absorption, molecules are transported via the portal vein to the liver where they might be subjected to hepatic metabolism. Metabolism can be pronounced to such an extent that a drug can be completely withdrawn from circulation by the first liver passage. For compounds undergoing extensive hepatic first pass metabolism, other administration strategies have to be found.[1]

To bypass intestinal absorption and hepatic metabolism, drugs could be applied intravenously. Intravenous application (i.v.) has the advantage to make drugs immediately available for distribution as they reach circulation without prior hepatic metabolism. Other invasive methods comprise sub- or intracutaneous application.[2] However, injections are associated with a certain infection risk and are generally not favored for self-application by a patient.

[1] One could think of pro-drug administration, where the active drug component becomes available just after being metabolized in the liver. However, this strategy requires a functional liver parenchyma.

[2] A major drawback is related to the varying constitution of the subcutaneous tissue depending on the body part. Varying blood flow rate and subcutaneous fat content can substantially alter drug kinetics.

A more elegant and non-invasive way to avoid first pass metabolism is the application over the mucosal tissue. Drugs diffuse passively into the submucosal capillaries and into venous circulation. Determinants for passive diffusion are molecular size, lipophilicity, and charge. Besides the buccal mucosa, other mucosal tissues can be used for drug application as well. Nasal and rectal applications are available for many drugs.[11]

Although the drug application via the skin seems at first glance very attractive, it is hampered to a certain extent by the physiological function of the epidermis, which is to safeguard the body from environmental impacts. To reach dermal microcirculation a drug has to diffuse through the numerous layers of epidermis.[3] As a result transdermal drug delivery can be delayed and prolonged, and is sometimes hard to control. Typical application domains are treatment of ischemic heart disease (nitroglycerine patch) and acute and chronic pain (opioid patches, like buprenorphine or fentanyl patches).[12, 13]

In cases where the pharmacological target is hardly reachable from the circulation (e.g., due to barriers) one could consider direct application into the target organ, surgically or by injection.[4] However, this administration route does usually not qualify for self-application. Trained staff and a medical facility are needed for safe administration. Thus, a single dosing becomes much more tedious and expensive than an oral formulation would and limits the drugs application range dramatically. Moreover, there are targets that do not qualify for direct application.[5] This is especially the case when strongly invasive surgical procedures would be needed and the potential risk of infections demands for an exhaustive risk-benefit assessment. In these cases, scientific ingenuity is needed to improve pharmacokinetic properties to qualify for safer application routes. However, one of the greatest obstacles is to overcome pharmacological barriers.

3.2.1 Barriers

Where body compartments are more sensitive to fluctuations of nutrients or exposure to xenobiotics, they need the ability to control and influence passage of molecules from circulation. Highly specialized cells fulfill this task by establishing biological barriers. Molecular trafficking can be controlled by active transport, often in combination with an increased tightness of the cellular layer, where the whole process is catalyzed. The characteristics of these barriers depend substantially on their location and the physiological requirements of the protected organ. We will discuss some of these barriers in the following in more detail.

[3] Of which, the stratum corneum imposes the major diffusion barrier as it mainly consists of several layers of dead ceratinocytes. Compounds would have to diffuse intercellularily through this inert barrier. In comparison, diffusion over the stratum lucidum, granulosum, spinosum, and germinativum of the epidermis is much faster due to a higher fluid content in these living cells.

[4] For the anti-angiogenetic agent Pegaptanib used in age-related wet macular degeneration the intravitreal injection is the common and most effective application route.[14]

[5] E.g., the brain.

3.2.1.1 Intestinal absorption

One of the first hurdles an orally administered compound encounters after dissolution in the gut is its intestinal absorption, i.e. its passage from the gut lumen into the portal vein. The cellular barrier in the GIT is mediated by a simple columnar epithelium of enterocytes. Current understanding indicates that passive diffusion (transcellular and paracellular) is a determining factor in drug absorption.[15]

The cellular membrane of the enterocyte consists of a self-assembling phospholipid bilayer. The aliphatic parts are oriented towards the inside, while the polar phosphate and head groups are directed toward the watery surrounding (e.g., cytosol and gut lumen). Before a compound moves by Brownian motion through the membrane it has to withdraw hydrating water molecules and to brake up hydrogen bonds. Generally, the higher the hydrogen bonding capacity, the more energy the permeation will cost and consecutively the poorer is the molecule's absorption. Due to this energy-consuming step, lipophilic and uncharged compounds permeate much better than their polar counterparts.[16, 17] Molecules passing through the polar head groups of the phospholipids encounter tightly packed lipid chains in the glycerol backbone. Hence, small molecules pass this region more readily than greater structures. Typically, measures of lipophilicity (polar surface area [PSA], partition coefficient [LogP]), size (molecular weight), and charge (hydrogen bonding capacity, PSA) are used to predict intestinal absorption in rules of thumb.[16, 18] The majority of molecules diffusing passively will take the transcellular route due to the great exchange area on the microvilli. But also paracellular diffusion occurs.[19, 20] Tight junctions between enterocytes control this undirected transport by claudine-pores, which act like a molecular sieve. Only small molecules (180-200kD) and mostly cations are able to cross. [21]

However, many vital substances are neither lipophilic, nor small (e.g., sugars and proteins) and will not diffuse passively in efficient manner through the enterocyte membrane. To ensure sufficient supply of such poorly permeable yet indispensable molecules, selective transport is warranted by several transmembrane transport proteins and channels. Beside specific import of molecules, there exists as well active extrusion of potentially noxious substances on the enterocyte. As they can transport their substrates against a concentration gradient, efflux transporters can modify absorption considerably. In enterocytes, transporters are physiologically involved in absorptive uptake (from the gastric lumen through the epithelial cells into the blood), in efflux (from the epithelial cell membrane back into the gastric lumen), and in secretory efflux (from the blood into the gastric lumen). [6]

Active influx and efflux at the level of the enterocyte are regulated by several transport systems, such as the influx transporter PEPT1 (Section 3.2.3.2) and the well-known efflux transporter P-glycoprotein (P-gp) (Figure 1) (Section 3.2.3.1).[23-25]

[6] Digoxin is secreted by P-gp form blood into the gastric lumen.[22]

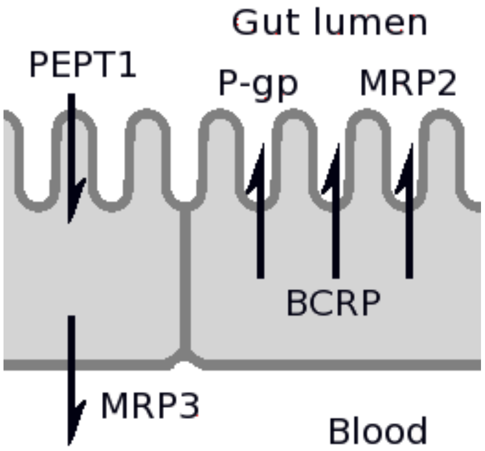


Figure 1 - A schematic view of enterocytes is given. On the apical side (gut lumen) P-glycoprotein (P-gp), breast cancer resistance protein (BCRP) and multidrug resistance protein 2 (MRP2) mediate efflux. Influx transporter peptide transporter 1 (PEPT1) mediates di- and tripeptide uptake. On the basal side (blood) multidrug resistance protein 3 (MRP3) transports substrates into the blood.

There is clear evidence that transport proteins interplay with metabolic enzymes. The effect of enterocytic cytochrome P450 (CYP450) metabolism, even though small when compared to the effect of hepatic CYP450, still serves as an example of metabolic degradation of the parent substance resulting in lower plasma levels.[26, 27]

3.2.1.2 Blood brain barrier

Development of new CNS active drugs is hampered by limited brain permeation. As invasive methods have proven themselves to be ineffective and risky for patients, the systemic application is the preferred route for drug administration into the brain.[28, 29] Hence, blood brain barrier (BBB) permeability is a feature absolutely mandatory for any drug, which targets the CNS. It is desirable to have estimates on a compounds behavior at the BBB as early as possible in the drug development process.

The microvascular endothelial cells of the brain establish the BBB. The membrane of brain endothelial cells exhibit negatively charged polar head groups, which oppose acids.[17][7] Circumferential tight-junctions connecting adjacent cells eliminate paracellular leakage and seal the physical barrier against paracellular diffusion of blood borne molecules (Figure 2). Lack of endothelial fenestration enforces the cellular barrier additionally.

[7] Acids penetrate poorly the BBB due to the negatively charged head groups of the lipid bilayer. This is also reflected in the fact that approximately 75% of the most prescribed drugs are basic, 19% are neutral, and only 6% are acids.[30]

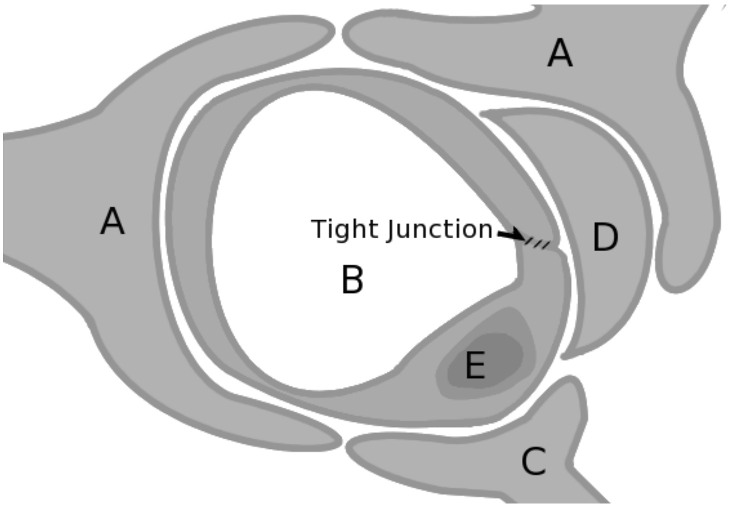


Figure 2 - A schematic intersection of a cerebral microvessel is shown. The microvascular endothelial cell (E) constitutes the blood brain barrier, which controls passage of molecules from the blood (B) into the brain. Tight junctions establish high intercellular resistance. The brain microvascular endothelial cells stand in close contact to astrocytes (A), neurons (C) and pericytes (D), which are thought to modify endothelial cell characteristics.

Therefore, most compounds have to take the transcellular route in order to cross the BBB. Small gaseous molecules (e.g., O_2, CO_2) and small lipophilic agents (e.g., ethanol) cross the endothelial cell membrane by passive diffusion.[31][8]

The process of passive permeation is well described and major physicochemical determinants summarized in rules of thumb, which are lipophilicity,[9] molecular weight, and measures of molecular polarity.[32-36] However, such expert-based rules do not accurately reflect the complexity of interactions as they disregard the pharmacokinetic processes mediated by transport proteins.[37] Typically, several anti-cancer drugs, corticosteroids, and anti-epileptics are well-documented examples where molecular properties for brain penetration would seem to be fulfilled but in fact significantly lower CNS concentrations are achieved due to their susceptibility to active transport.[38, 39] Physiologically, the ABC transporter super family and solute carriers mediate active transport across the BBB and constitute a biochemical barrier to safeguard the brain tissue from potentially toxic compounds, such as xenobiotics.

[8] High lipophilicity improves brain permeation, which is nicely demonstrated on the example of morphine: Addition of methyl groups to morphine produces codein, which penetrates 10 fold better into the brain. When two acetyl groups are added, which make the compound even more lipophilic, heroin is produced which further increases permeability (up to 100 fold).

[9] However exaggerated lipophilicity makes a compound susceptible for nonspecific binding. It is therefore important to balance lipophilicity in order to achieve optimal pharmacokinetics.

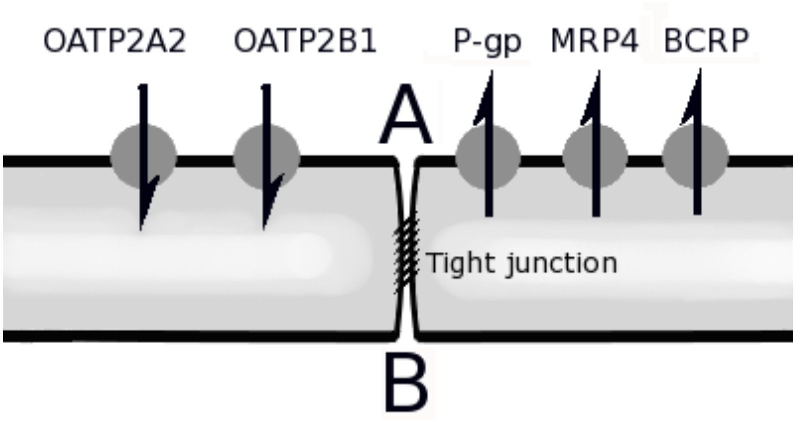


Figure 3 - An intersection of microvascular endothelial cells is shown. Influx is mainly mediated by organic anion transporting polypeptides OATP2A2 and OATP2B2. P-glycoprotein (P-gp), breast cancer resistance protein (BCRP) and multi drug resistance protein 4 (MRP4) are examples of efflux transporters.

P-glycoprotein (P-gp, ABCB1) and breast cancer resistance protein (BCRP, ABCG2) are the most prominent and best characterized representatives and show the highest mRNA expression levels of all ABC transporters on the human BBB.[40-44] Their impact on substrate drug uptake has been shown to be, at least for P-gp, clinically relevant.[45] There are speculations that both transporters act together to prevent brain entry of several toxic compounds (Section 3.2.3).[46-48] Only a very small proportion of compounds show enhanced permeation due to uptake transporters (Figure 3). The physiological role of these transporters is uptake of nutrients like sugars, peptides, amino acids, and other endogenous compounds (Section 3.2.3).[49, 50]

In the past, the most commonly used brain penetration data were derived from *in vivo* pharmacokinetic studies, which produced a drug in brain to drug in plasma/blood ratio at steady state. Usually its logarithm was used termed LogBB. This measure can give some indication of distribution in the brain, however it suffers from limitations.[10] Single time point measurements might not accurately reflect brain penetration due to varying kinetics in plasma and brain. Moreover, LogBB reflects a volume of distribution that is determined largely by cytoplasmic binding of drugs in brain and much less by BBB permeability. This measurement cannot resolve whether the fraction of free drug is camouflaged by nonspecific or specific binding nor does it provide any information on active transport.[51] Therefore, the permeability surface product values are recommended, which are usually calculated from internal carotid artery perfusion studies in rats, given as its logarithm, LogPS. This procedure is considered

[10] The term was loosely applied for variously calculated data: LogBB was sometimes derived from area under thr curve (AUC) values, steady state or single time point measurements. In order to make use of these values, the scientist had to have knowledge on how the data were derived.

superior to blood/brain partitioning measurements at steady state, as it lacks systemic distribution effects, which distort brain penetration substantially.[51]

3.2.1.3 Blood milk barrier

To date, experts estimate a nursing rate of 60–90% in western countries[11] and breastfeeding is considered the best nutrition for the first months of a baby's life.[52-56] While in general mother and baby profit from nursing, maternal medication intake can impose a safety concern. As many drugs pass easily into breast milk, babies can be accidentally exposed to medication. Although the majority of drugs do not impose a hazard, some cases of significant infant intoxication exist.[57]

Almost all lactating women receive some medication immediately postpartum and during nursing.[58][12] Despite its social, economic, and medical impact, compatibility of drug intake in nursing is still a relatively unexplored field. Ethical constraints hamper clinical trials and animal tests give only a rough estimate of human pharmacokinetics. As a consequence, for many drugs only case reports exist.[13]

Passive diffusion is a leading mechanism of drug passage into breast milk.[61, 62] To our knowledge, highly passive diffusing molecules are determined by factors such as low molecular weight, high lipophilicity, and low polarity (Section 3.2.1.1).[63] Pharmacokinetics and plasma protein binding in maternal circulation determines the amount of drug, which becomes available for excretion.

Although excretion into breast milk is predominantly guided by passive diffusion, the occurrence of drug accumulation in human and animal milk suggests the presence of active transport in the mammary gland.[61, 64-66] The lactating mammary gland epithelium has to secrete vitamins and nutrients against a concentration gradient. Coherently, a multitude of transport proteins were found to be expressed.[67, 68]

Members of the ABC transport protein family, like breast cancer resistance protein (BCRP, ABCG2), are expressed on the mammary gland epithelium.[64-66] Surprisingly, in the lactating breast, BCRP concentrates drugs, carcinogens, and toxins into milk.[69, 70] This behavior stands in sharp contrast to its detoxifying function in other organs, for example in the placenta, where it transports noxious substances against a concentration gradient from fetal to maternal circulation (Section 3.2.1.4). Herwaarden and co-worker suspected that toxin accumulation in breast milk is most likely due to a usurped physiological mechanism. BCRP might serve to concentrate vitamins and nutrients in breast milk as secretion of Riboflavin (Vitamin B2) by BCRP has been shown.[71]

[11] The nursing rate in developing countries is presumably even much higher.

[12] An increased vulnerability to psychiatric conditions (e.g., depression)[59] and treatment re-uptake after pregnancy leads to a high incidence of drug prescriptions in breastfeeding mothers.

[13] Consequently, manufacturers' information on drugs is often overly cautious due to lacking experimental experience. Hence, mothers are often advised to stop nursing rather than to risk drug exposure for the baby.[60]

The probability of adverse events from accidental drug intake via maternal milk might rise with increasing exposure (e.g., accumulation), but toxicity of a compound also depends significantly on drug clearance of the infant. To link milk plasma/serum ratios (MP) with infant drug clearance and milk intake a simplified "Exposure Index" has been proposed by Ito and co-workers:

$$\text{Exposure index} = \text{A} \times (\text{M/P ratio})/\text{Clearance}$$

where *A* is a coefficient (10ml/kg/min), *M/P ratio* is the milk plasma ratio, and *Clearance* is the drug clearance of the infant expressed as ml / kg / min. [72] Infantile drug clearance depends highly on renal and hepatic metabolism and excretion. Characteristically, the glomerular filtration rate of a newborn achieves adult values 3-5 months after birth, while tubular secretion rate matures more slowly, accounting for prolonged elimination half-lives.[73] Expression of drug efflux transporters on liver and gut wall, such as P-gp and BCRP might be highly subjected to individual development.[74] Estimating drug clearance in infants is therefore a difficult undertaking.

3.2.1.4 Placenta barrier

The physiological function of the placenta is the exchange of gas, import of nutrients as well as export of fetal waste products. Moreover, it has a protective function as it saves the fetus from toxic compounds from maternal circulation. In contrast to former beliefs, the placenta barrier does not mandatorily protect against harmful drug exposure, as the Thalidomide scandal in the 1950s impressively demonstrated.[75]

In the placenta, the main diffusion barrier is mediated by the fetal syncytiotrophoblasts, which directly invade the uterine wall. The predominant mechanism of molecule exchange is transcellular diffusion (Section 3.2.1.1). Active transport mechanisms support passive permeation of glucose, peptides, and other vital molecules (Section 3.2.3). ABC transporters like BCRP and P-gp are strongly expressed and mediate efflux on the syncytiotrophoblasts.[76]

3.2.2 Metabolism

Once absorbed, drugs are transported via the portal venous system into the liver, where hepatocytes absorb and modify molecules to increase water-solubility. Drug uptake on level of hepatocytes happens mostly against concentration gradients and is facilitated by a multitude of transport proteins (Section 3.2.3). After modification, drugs are either eliminated (via bile) or re-circulate into systemic blood flow and are distributed in the body.

The compulsory shunting of intestinally absorbed molecules to the liver accomplishes two important tasks. Nutrients, such as fats or sugars, are modified and/or stored and noxious substances can be removed from circulation before they are distributed in the body. The liver exhibits the capability of eliminating drugs completely in the first passage from portal venous circulation. Hence, hepatic first pass metabolism can influence bioavailability considerably.

Metabolic modification can also lead to activation of drugs. This principle is exploited by pro-drugs. While the administered compound is inactive, the drug is activated by biotransformation in the liver. This strategy was applied to improve absorption of the drug oseltamivir, where the active ingredient (oseltamivir carboxylate) exhibits poor intestinal absorption capacity. By methylation the drug becomes absorbable orally and is almost completely hydroxylized in the liver to its active component.[77] Pro-drugs can also be used to enforce oral application, e.g., to avoid i.v. drug abuse.[14]

To reach the site of metabolism substances have to be efficiently transported into the hepatocyte. Principally, the same active and passive transport mechanisms are involved as in enterocytes (Figure 4).

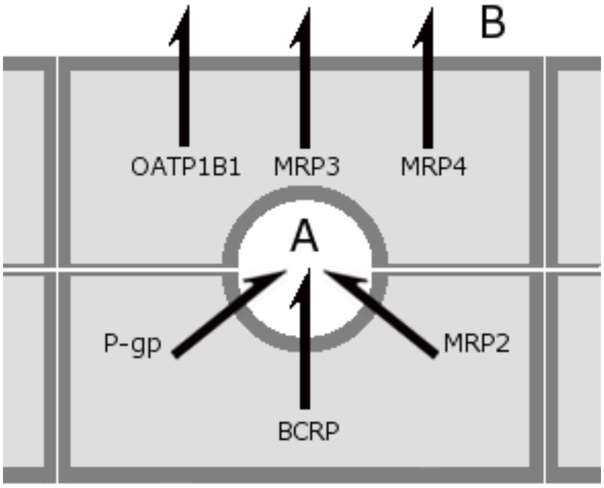


Figure 4 - Schematic view of transport proteins on hepatocytes. Organic anion transporting polypeptide OATP1B1 mediates influx of substrates into the cells. After metabolic modification, compounds are either excreted apically into the bile canaliculi (A) or transported back into circulation (B) for renal excretion or/and systemic distribution. In the hepatocyte, P-glycoprotein (P-gp), breast cancer resistance protein (BCRP) and multidrug resistance protein 2 (MRP2) mediate apical export. Multidrug resistance protein 3 (MRP3) and 4 (MRP4) transport substrates back into the blood flow.

Transport proteins are of particular importance in hepatic clearance. They enhance biotransformation by facilitating uptake into hepatocytes, where molecules encounter metabolizing enzymes (Section 3.2.2). They also mediate clearance by increasing the efflux of metabolites into the bile canaliculi or back into the blood stream. Single transport proteins are discussed in Section 3.2.3.

[14] Valorone N is a mixture of the opiate tilidine and the opioid antagonist naloxone. It is claimed that due to naloxone's high first pass metabolism, oral administration is mandatory to experience a pharmacological effect of tilidine. When applied intravenously, naloxon becomes systemically available and antagonizes the effects of tilidine.

3.2.2.1 Sites of metabolism

Although, the majority part of metabolism takes place in the liver, metabolic enzymes are practically ubiquitarily expressed and contribute substantially to modification and excretion of nutrients and xenobiotics.

Intestinal metabolism can affect drug absorption. On the other hand, several drugs and nutrients (e.g., green tea extract or hypericum) can induce intestinal metabolic enzymes, such as CYP 450. [78, 79] In the brain, glial cells and neurons express metabolic enzymes and there is further evidence that also brain endothelial cells have a metabolic function, at least in disease.[80-83] The list could be continuously elongated. However, the wide spread presence of metabolic enzymes underlines their impact on both the maintenance of homeostasis and also drug excretion.

3.2.2.2 Molecular mechanisms of metabolism

Metabolism is usually a two-step process, which has not necessarily to occur in sequence. The first reaction is characterized by modification of molecular structures by oxidation, hydroxylation, or reduction. Step two-reactions are usually additions (conjugations) of polar groups, such as glucuronic acid, amino acids, or glutathione, which increase hydrophilicity. A compound does not mandatorily need to undergo step one before step two, if it already has a functional group qualifying for conjugation.

The most prominent phase I enzymes are monooxygenases which include the CYP450 family. They are localized on the endoplasmic reticulum and abundantly expressed in hepatocytes. CYPs are also found in the intestine, colon, lung, brain, and skin.[84, 85] Several members of the CYP protein family show polymorphisms, which led to unexpected pharmacokinetics of substrate drugs in certain populations.[86, 87] Numerous drugs and herbal preparations are inducers of CYP and complicate drug therapy considerably. [88]

Uridine 5'-diphospho-glucuronosyltransferases (UGT)[15] play the predominant role in phase two of metabolism. Substrate molecules are conjugated to either a glucuronic acid moiety, a hydroxyl carboxylic acid or an amine group. Glucuronisation increases water solubility and hence eases renal and biliary elimination. Some hereditary diseases are connected with UGT abnormalities or deficiencies, such as Gilbert-Meulengracht Syndrome[16] and Crigler-Najjar[17] Syndrome.

[15] UGT is expressed practically in all animals and plants, except in cats (*genus felis)*, where it accounts for a series of unusual toxicities.[89]

[16] Gilbert-Meulengracht Syndrome is characterized by a mild hyperbilirubinemia and is found in approx. 5% of the population. The disease is caused by a reduced activity of UGT1A1. Substrate drugs show an increased toxicity in these patients (e.g., Irinotecan). However phenobarbital can induce and restore activity of UGT1A1.

[17] Crigler-Najjar Syndrome is a very rare autosomal recessive disease. Type 1 is characterized by sever non-hemolytic hyperbilirubinemia caused by a complete lack of UGT1A1. Untreated, the hyperbilirbunemia leads to severe brain damage or even death. In Type 2, disease is less severe, as UGT1A1 expression is reduced and not completely abolished.

Besides UGT, sulfotransferase, glutathion-S-transferase, and N-acetyltransferase catalyze phase II metabolism, conjugating sulfate groups, glutathion, and amines, respectively. After conjugation, compounds are subjected either to excretion in the bile or to recirculation into the systemic blood flow for renal clearance.

3.2.3 Active Transport across membranes

Transport proteins have specified substrates and exploit individual transport mechanisms. Some transport their substrates along the concentration gradient (facilitated diffusion) while others use energy to overcome this gradient actively.[90] In contrast to passive diffusion, these active transport processes exhibit saturation kinetics (Figure 5).

Principally speaking, transport occurs when the drug contains a moiety that is similar to transporters natural substrate or if it has structural elements that facilitate binding to the transport protein (e.g., P-gp). Transporters affect absorption, distribution and toxicity properties in various ways, which have to be considered in drug development.

For certain drugs, an enhanced intestinal absorption can be observed, despite unfavorable physicochemical properties.[50] Examples of these drugs are peptidomimetics, like beta-lactam antibiotics or ACE inhibitors, and anti-viral, and anti-cancer, drugs which are transported via PEPT1. [91-93] Inversely, some molecules are badly or not absorbed (e.g., anti-cancer drugs) due to efflux transporters.[94] They can oppose distribution or enhance elimination. Competitive inhibition as well as induction of transporters can additionally modify pharmacokinetics.[95]

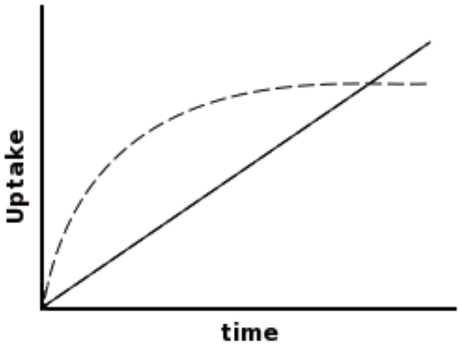


Figure 5 - Diagram of active (dashed line) and passive transport (continuous line) kinetics. Active transport is characterized by an increased uptake until all transporters operate at full capacity, i.e. are saturated. Transport rates are stabilized regardless of excess substrate. Passive diffusion shows a linear kinetics, which continuously increases with increasing concentrations.

Owing to the finite number of transport proteins on the cell surface, active transport can be saturated if substrate is available in sufficiently high concentrations. The flux of molecules increases until the maximum capacity of the transport proteins is reached. Above this level the flux does not increase. This effect is not seen in passive diffusion, which exhibits linear and not saturation kinetics (Figure 5).[18] Transporters are found at barrier membranes throughout the body. Some of the most important ones shall be discussed in more detail.

3.2.3.1 ABC transport proteins

P-Glycoprotein (P-gp, ABCB1) is probably the best-characterized member of the ABC transporter super family. It is an ATP-dependent drug efflux pump exhibiting broad substrate specificity.[96, 97] P-gp exhibits 12 trans-membrane domains (Figure 6). To undergo transport, substrates have to attach to the binding domains of P-gp, of which one appears to be within the cellular membrane. By hydrolyzation of two ATP molecules on the ATP binding regions, P-gp changes conformation, opening a pathway for the substrate to be extruded into the extracellular fluid.[90, 98]

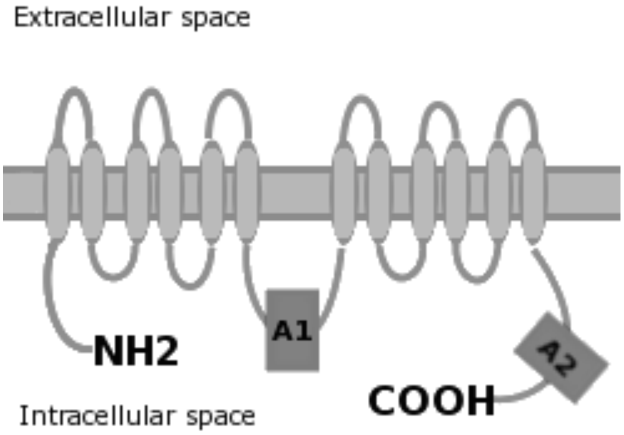


Figure 6 - Schematic view of P-glycoprotein, with typical 12 trans-membrane domains. The ATP-binding sites are indicated by dark grey boxes (A1 and A2).

P-gp was discovered, as decreased drug concentrations and a consecutive multidrug-resistance in tumor cells was observed.[96] It has a protective and excretory function in physiological tissues, and is abundantly expressed on several barriers. Thierbaut and co-workers demonstrated the expression of the transporter on the apical side of enterocytes, hepatocytes, brain endothelial cells, and the proximal tubule of the kidney. [99]

P-gp exhibits broad substrate specificity and can substantially influence pharmacokinetics in clinically

[18] This holds in the case of stable concentration gradients.

relevant manner.[19] In the intestine it reduces or abolishes uptake of substrates, whereas it enhances in the liver and kidney clearance of substrates into bile and urine, respectively.[90] Schinkel and co-workers reconfirmed its relevance in drug transport as they found significantly elevated substrate levels in P-gp deficient mice. [98, 102, 103][20]

P-gp plays a detrimental role in supporting cellular barriers such as the BBB or placenta barrier. Recent reports even indicate P-gp expression on the mammary gland epithelium. [104, 105] The intentional application of P-gp inhibitors as a "chemo-sensitizer" in order to enhance efficacy of drugs or to reduce the active component in a single dose was recently discussed.[106-108] Although tempting, this approach might harbor risks. Inhibition of this transporter, which is expressed in many tissues might corrupt its protective function in other organs as well, e.g., the BBB (Section 3.2.1.2), potentially leading to acute intoxication by overdose.

Breast cancer resistance protein (BRCP, ABCG2) was identified from chemotherapeutically resistant breast cancer cells.[109] It is a "half" ABC transporter as it exhibits only six trans-membrane domains.[110] Physiologically, BCRP shows high expression levels in the gastro intestinal tract, liver, kidney, brain endothelium, mammary gland and the reproductive organs.[111] Physiologically, BCRP contributes to efflux of porphyrines and shares many substrates with P-gp. As both transporters are often found in co-localization and show a broad substrate overlap, it was suspected that they work in a concerted manner. Gastrointestinally expressed, it limits absorption of its substrates, such as sulfasalazil.[112] In reproductive organs, BCRP safeguards sensitive tissues from noxious agents. Additionally, the transporter is found on the apical membrane of the hepatocytes, where it mediates together with P-gp and MRP2 excretion.[90] BCRP substrates comprise antiviral drugs (e.g., zidovudine), statins (e.g., rosuvastatin), antibiotics (e.g., ciprofloxacin), and calcium channel blockers (e.g., azidopine).[111, 113] The extraordinary role of BCRP in the lactating breast is discussed in Section 3.2.1.3.

The multidrug resistance proteins (MRPs, ABCC family) share less than 15% amino acid identity with other members of the ABC transport protein family. The similarity resides almost exclusively with the nucleotide biding domains. MRPs are primary active transporters and mediate the ATP-dependent unidirectional transport of lipophilic substances conjugated with glutathione, glucuronate, or sulfate and conjugated and unconjugated amphiphilic anions. The expression of MRPs was first described in the doxorubicin selected lung cancer cell line H69AR, which showed resistance to many chemotherapeutic agents.[114] Their expression was thereafter confirmed for a broad range of human tumors and various healthy human tissues.[115] The family of the MRPs consists of at least six members, of which MRP3 (ABCC3) and MRP4 (ABCC4) have a certain role in disposition, and the apically localized

[19] When rifampicin, a potent inducer of P-gp, is coadministered with digoxin, a P-gp substrate, the absorption ratio of digoxin was significantly reduced.[100] The inverse effect was observed when quinidin, an inhibitor, is given instead of rifampicin. Serum levels of digoxin increased up to two- and threefold in healthy subjects.[101]

[20] In fact, serum levels of P-gp substrate ivermectin were 20 times higher than in wild type animals.

MRP2 (ABCC2; also known as cMRP or cMOAT) is suspected to have emerging clinical importance.[90, 116, 117] MRP2 is strongly expressed on the apical canalicular membrane of hepatocytes, where it promotes biliary efflux of glucuronides, sulfates, glutathion, and amphiphilic organic anions.[118] However, MRP2 is also found on apical membranes, on the proximal tubules of the kidney, in the intestine, as well as on the placenta, and in the lung.[119-121] It is often co-localized with phase two metabolism enzymes (e.g., UGT), which produce some of MRP2's substrates.[122] Moreover, it was shown that vectorial transport in MRP2 transfected cells happened only in presence of influx transporters, such OATPs, which indicates that MRP2 might mediate drug interaction in coordination with influx transporters and metabolism.[123-125] MRP2's presence on the human BBB is debated.[90] However, its overexpression was associated with phenytoin resistant epilepsy in rats.[126, 127]

Physiologically, the transport protein plays an important role, as its localization on many apical membranes (e.g., in the liver or kidney) makes MRP2 the final elimination step for many drugs and xenobiotics.[128] Dysfunctional expression or inhibition of MRP2 can results in unusual toxicities, like the conjugated hyperbilirubinemia in Dubin-Johnson syndrome.[129][21] MRP2 can alter pharmacokinetic properties of anti-cancer drugs (e.g., methotrexate and mitoxantrone), antibiotics (e.g., ampicillin and rifampicin), angiotensin receptor antagonists (e.g., valsartan and olmesartan).[90, 131, 132] The exact substrate binding sites and mechanisms leading to induction and inhibition are not yet completely elucidated.[133] Moreover, the controversial role of glutathion as transport stimulator and co-transported agent indicates the complexity of the process.[134, 135]

3.2.3.2 PEPT1/2

The tertiary active peptide influx transporter PEPT is expressed in two isoforms, PEPT1 and PEPT2. Both are expressed on the proximal tubule of the kidney, while PEPT1 is exclusively found on the apical membrane of enterocytes.[90] It typically recognizes di- and tripeptides, but not individual amino acids. Peptides are internalized against a concentration gradient in co-transport with a proton. In order to keep the intracellular proton concentration low, the Na^+/H^+-Exchanger protein 3 extrudes protons on the apical side in exchange with Na^+-ions. A basolaterally located Na^+/K^+-ATPase maintains intracellular Na^+ ion concentrations. PEPT transports not only peptides but also drugs, which resemble peptides. Peptide-like drugs, like beta-lactam antibiotics and ACE inhibitors are absorbed in higher concentrations, as their physicochemical properties would let expect.[91, 92] [22]

[21] The autosomal recessive Dubin-Johnson Syndrome exhibits a deficiency for MRP2 and is characterized by intermitting hyperbilirubinaemia. Though it seems that MRP3 may rescue the export of conjugates across the basolateral membrane. This was also reported for other conditions where the canalicular secretion of MRP2 substrates is impaired.[130].

[22] To improve unfavorable intestinal absorption, drugs can be linked to an amino acid rest to resemble peptide structure and become PEPT1 substrates. The pro-drug valacylovir achieved 50% better absorption ratios by conjunction to valin than to its un-linked parent compound acyclovir.[136] Other successfully modified drugs are L-dopa (L-Dopa-L-Phe) and gangcyclovir (valgangyclovir).[137]

3.2.3.3 OATP

Organic anion transporting polypeptides (OATP) are a family of influx transporters, which physiologically import conjugated and unconjugated bilirubin, bile acids, conjugated steroids, and thyroid hormones.[138-140] OATP1B1, OATP1B3, OATP2B1 are mainly expressed on the sinusoidal membrane of hepatocytes where they mediate substrate influx from the blood flow. [141]

OATP1B1 is probably the best-characterized member of this family. It is difficult to estimate the role of OATP1B1 in drug-drug interaction in isolation as OATPs share many substrates with other transport proteins (e.g., MRP2) and metabolic enzymes.[142] However, several drugs are known to be transported by OATP1B1, such as statins, ACE inhibitors, and angiotensin II receptor antagonists. A typical substrate often used in experimental settings is the antihistamine fexofenadine.[141] A typical inhibitor of OATP1B1 is cyclosporine as its coadministration lead to increased statin levels.[143] OATP polymorphisms can cause marked differences in pharmacokinetics. A polymorphism of OATP1B1 lead to reduced substrate specificity of simvastatin, which increased the risk of drug induced myopathy.[144][23]

OATP1A2 is mostly located at the luminal membrane of small intestine and the BBB.[146] Its physiological and drug substrates resemble those of OATP1B1. Its uptake function can be inhibited by naringin found in grapefruit and orange juice.[147, 148]

3.2.4 Ways to assess pharmacokinetics

Assessment of pharmacokinetics is a complex and difficult endeavor as it becomes clear from the multitude of processes involved outlined above. *In vitro* models can give information on pharmacokinetics on the cellular level. In order to assess pharmacokinetics as a more realistic multistep process, animal models are usually needed, however, this has the significant disadvantage that testing compounds *in vivo* is expensive and time consuming. As nowadays, pharmaceutical companies harbor chemical libraries of millions of molecules, computational methods pose an economic and efficient alternative to screen for potential lead compounds. *In silico* methods can compress immense quantities of information in predictive models. By the mathematical projections of molecules they can reveal new mechanistic explanations of the process itself.

[23] Interestingly, fluvastatin seems not to be affected by this polymorphism.[145]

3.3 QSAR, Quantitative Structure Activity Relationship

3.3.1 Fundamentals

The first attempts to relate chemical structure and biological action were taken in the mid-19th century in the field of toxicology. In 1863, Cros stated in his thesis a relationship of toxicity and water solubility of primary aliphatic alcohols.[149] He related pharmacological behavior to molecular properties, which in turn were determined by a compounds structure. Crum-Brown and Fraser refined this observation.[150] They stated that the physiological action of a molecule in a certain biological system (Φ) is a function (f) of its chemical constitution (C):

$$\Phi = f(C)$$

From this, they deduced, that an alteration in chemical constitution (ΔC) would be reflected in a change of biological activity ($\Delta\Phi$).

3.3.1.1 Similarity principle

A fundamental prerequisite for QSAR was the formulation of the similarity principle. It relates chemical structure to functional behavior, stating similar structures exhibit similar activity. In 1874, Körner proposed the first correlations between molecular structures and physicochemical properties.[151, 152] His work dealt with the ortho-, meta-, and para- derivatives of benzene. The different colors of the derivatives were related to the differences in chemical structure. The indication of ortho-, meta-, or para-substitution can be seen as the first molecular descriptor.[24]

First quantitative property-activity studies (QSPR) in classical meaning where published in 1893 by Charles Richet. He correlated water solubility of ethanol, diethyl ether, urethan, paraldehyde, amyl alcohol and absinth extract with their lethal doses in dogs.[154] He stated, "*plus ils sont solubles, moins ils sont toxiques*", the more water soluble, the less toxic compounds are. This was the first inverse linear relationship formulated of solubility and biological activity.

At the turn of the century, several works correlated narcotic drug potential to water/oil partition coefficients, to molecular chain length, or to surface tension.[155-157][25] Louis Plack Hammett compared in 1938 dissociation rates of different benzoic acid derivatives with meta- and para-

[24] A decade later, Mills found a relationship between structure and melting and boiling point of a homologous series of compounds.[153]

[25] Overton positively correlated narcotic potential of drugs with their solubility in olive oil. His observations were independently reconfirmed by Meyer and were put forward as the Meyer-Overton hypothesis. However, the thereof resulting lipid theories cannot explain receptor-mediated reactions.

substituents.[158] He observed that similar substitutions on different aromatic compounds resulted similar effects, which led him to deduced the seminal Hammett equation, which states

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

where ρ is the reaction constant, depending solely on the reaction type, σ is the substituent constant depending on the substitute. K and K_0 are the dissociation constants of two distinct molecules. In other words, the reaction depends solely on the reaction type and the substitute group.

In the end of the 1940s, the first relationships of biological activities to theoretical numerical indices were drawn. Examples are the Wiener Index and the Platt Number derived from graph theory (Section 4.2.3).[159-162] In the following decade, a multitude of features were derived from the graph theory, marking the beginning of systematic studies on molecular descriptors.

In mid-1960s, Hansch and co-workers gave the quantitative structure activity/property relationship (QSAR/QSPR) approach its modern face, by publishing their pioneering work on structure activity relationships in plant growth regulators and their dependency on Hammett constants and hydrophobicity.[163] They determined a series of octanol-water partition coefficients (LogP) and introduced a new hydrophobic scale to characterize permeation of molecules through hydrophilic environments, such as blood or membranes.

3.3.1.2 Dimensionality

Dimensionality of QSAR models usually refers to the techniques and descriptors used to create them. In the beginnings of QSAR, activity was related to experimentally assessable parameters and those deducible from chemical notation, i.e., physicochemical properties. These features are usually referred to as one-dimensional (1D) descriptors (e.g., molecular weight).

At the end of the 1960s, Free and Wilson proposed modeling biological responses on substitution effects on common molecular skeletons.[164] Additionally, introduction of graph theory lead to descriptors, which make statements on connectivity of molecules as a whole. The molecular graph is a two-dimensional representation of a compound and hence the thereof deduced descriptors are usually termed two-dimensional (2D) descriptors.

The consideration of actual spatial distribution and geometry of a molecule led to three-dimensional (3D) descriptors. These are typically charged partial surface area (cPSA) introduced by Stanton and Jurs, and gravitational indices by Katritzky and co-workers.[165, 166] It was debated that the connection table holds enough implicit sterical information that effective use of 3D coordinates would not add much more geometrical information.[167]

The introduction of induced fit modeling expanded the dimensionality to four-dimensional (4D) and even higher dimensional levels. Although higher dimensional models hold in general more information, it was argued that increasing dimensionality does not mandatorily yield superior models.[168] Generally, the model should be suited to reflect the underlying data and to meet the demands of their application. There are completely different requirements on a database screening compared to a single molecule analysis using pharmacophores and computing conformational changes. For these reasons, the question on superiority of models cannot be finally answered.

3.3.2 Applicability Domain

Whether a QSAR model can establish an accurate and reliable prediction of an unseen structure, is determined by its applicability domain. It is defined as the information space a model has been generated on. The accurateness of predictions is only warranted within the scope of its applicability domain and this holds usually true for interpolation rather than for extrapolation.

As no generally acknowledged measure for the applicability domain has jet been proposed, it is recommendable to describe it with the most relevant parameters, which are usually the (physicochemical) descriptors used to create the model. The molecules can be represented in the multi-dimensional space spanned by their descriptors and can be compared for structural similarity. However, the perception of similarity is subjective and a multitude of measures exist. Different endpoints require individual measures of similarity. Common similarity measures are summarized in Table 1.

Similarity measure	Notation
Tanimoto coefficient	$\frac{I}{A+B-I}$
Hamming distance	$A+B-2I$
Euclidian distance	$\sqrt{(A+B-2I)}$

Table 1 - Three similarity measures and their formulas are given. I is the intersection of the samples A and B.

4 Materials and methods

4.1 Molecular representation

One of the most fundamental prerequisites for computational chemistry is the formulation and definition of an accurate and unique chemical structure representation. The seminal idea of applying a graph theoretical approach to molecular representations revolutionized chemistry.[169] One could deduce from a chemical structure a two-dimensional hydrogen depleted molecular graph, where atoms and bonds are represented by vertices and edges, respectively. The simplest chemical graphs do not discriminate higher ordinal bonds or atom types. This made application of mathematical operations from the field of graph theory possible for molecules. A multitude of molecular descriptors (Section 4.2.3) and the SMILES representation for chemical structures (Section 4.1.1) are deduced from this pioneering idea.

4.1.1 Simplified Molecular Input Line Entry System (SMILES)

David Weininger defined in the late 1980s the SMILES concept that has become a standard through out computational chemistry. He proposed a string representation of molecules based on the chemical graph theory.[170, 171] Molecules are represented as ASCII string, which is human legible and easy to compute. While atoms are represented by their atomic symbols, branching points are indicated by parentheses and a numeric label indicates ring connection points. Lower case letters indicate aromaticity. Disconnected elements, such as salts, are indicated with a point (Table 2).

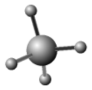
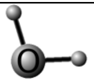
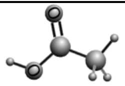
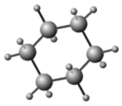
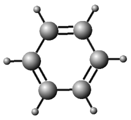
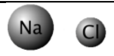
SMILES notation	Generic name	Structure
C	Methane	
O	Oxygen	
CC(=O)O	Acetic acid	
C1CCCCC1	Cyclohexane	
c1ccccc1	Benzene	
[Na+].[Cl-]	Sodium chloride	

Table 2 - Examples for SMILES notation, the corresponding generic names and three-dimensional structure are given.

If not specified, bonds are implicitly assumed to by single or aromatic bonds. Multiple bonds can be specified using the equality sign (=) and the number sign (#) for double and triple bonds, respectively. Table 3 gives an overview.

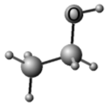
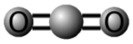
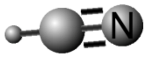
SMILES notation	Generic name	Structure
C-C-O (or simply CCO)	Ethanol	
O=C=O	Carbon dioxide	
C#N	Cyanide	

Table 3 - Bond notation of SMILES notation is illustrated.

4.1.2 SMILES Arbitrary Target Identification (SMARTS)

Typically, chemical databases are screened for similarity regarding molecular structure or activity. An efficient way to find resembling molecules would be a substructure search, i.e., the formulation of a subgraph of the molecular representation. The subgraph can then be used as a search pattern. SMART language is closely related to the SMILES code and allows for efficient search query definition.[172] SMARTS is built on the SMILES language, but is extended with logical operators and wild cards for bond or atom types. Some examples are given in Table 4. Equipped with these additional features, SMARTS language is a very efficient way to specify sensible search queries.

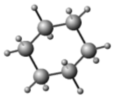
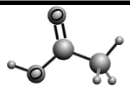
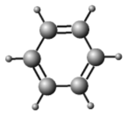
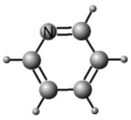
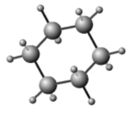
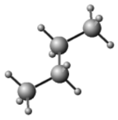
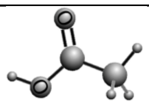
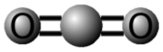
SMARTS	Short explanation	Example 1	Example 2
[#6]~[#6]	Two carbon atoms (atom No. six) connected by any bond		
[a]	Any aromatic atom		
[!c]	Not a aromatic carbon		
[O]~[*]	Any atom connected by any bond to oxygen		

Table 4 - A selection of SMARTS expressions and a short explanation is given. Additionally, two example structures are shown, which fulfill the corresponding SMARTS query.

4.1.3 Fingerprints

Fingerprints represent structural information on a compound as a feature vector and were initially intended for similarity or substructure search. The feature vector does not necessarily hold information in numerical form and can represent structural features also as a bit vector, where every bit codes for presence or absence of a chemical substructure. This is especially advantageous when large databases have to be screened for similarity. Once all fingerprints have been computed for a set of molecules, overlaps of bits can be compared in order to prescreen for similar fragments.

Which chemical substructures or properties are used to describe the molecules depends on the type of fingerprint used. Therefore, the vectors may also substantially vary in length. They range from 3D pharmacophore keys, which can be exceedingly extended, over fixed length 2D fingerprints, hashing connectivity patterns or chemical fragments to 1D features. Extended connectivity fingerprints were designed to explicitly consider features relevant for molecular activity and capture the local atomic neighborhood.[173] Another set of broadly used fingerprints are the fixed length 166 bit MACCS keys. Initially defined by the company MDL,[26] these fingerprints are most commonly used binary fragment-based keys.[174] All bits in the set represent a predefined set of chemical fragments (e.g., a aromatic ring structure or double bound oxygen) represented as SMARTS strings, which can either be present

[26] The company MDL (now known as Symyx, which has merged with the company Accelrys) also developed the MDL mol file, a standard chemical file format.

(bit set to ON) or not (bit set to OFF). In order to compare the ability of fingerprints to describe a dataset, one could use the so-called "fingerprint darkness". This concept refers to the fraction of bits set to ON in a binary vector. It is obvious that the more features a fingerprint captures the more discriminatory power it has.

4.2 Descriptors

Descriptors are probably best explained as mathematical representation of chemical properties. When applied in QSAR, they desirably help to identify molecules with comparable activity but at as high structural diversity as possible.[27] This conflicting situation might be a reason that we are in a continuous search for new molecular representations. In the following section, we will discuss some of the most widely used descriptors.

4.2.1 Constitutional descriptors

Constitutional descriptors are features that reflect molecular composition without any geometrical or topological information. In the early days of QSAR, some of these features were manually determined for a set of compounds but nowadays it is general practice to compute these features, which holds the advantage of not underlying experimental variation.

4.2.1.1 Atom count descriptors

Count descriptors are a relatively simple way to get information on molecular constitution. Atom numbers (or atom count) is the simplest measure for molecular size. Usually only non-hydrogen atoms are counted. The information index on size (I_{size}) can be derived thereof, which gives the total information content on atom counts. The formula is adapted from [175]

$$I_{size} = A \log_2 A$$

where the atom count *A* can also take hydrogen atoms into account, depending on its definition. Other count descriptors assess the contribution of heteroatoms (heteroatom count) or functional groups, like hydrogen bond donors and acceptors (Section 4.2.2.1).

4.2.1.2 Bond count descriptors

Bond number or edge counting refers to the simplest graph invariant of the molecular graph where also multiple bonds are considered as single edges. As a result it does not discriminate chemically non-equivalent groups. If information on molecular saturation is desired from the set of constitutional descriptors, double-, triple-, or aromatic-bond counts can be considered. Bakken and Jurs proposed the multiple carbon bond index to assess carbon bonds by their simple addition.[176] Another

[27] Favorably, descriptors should be easily computable and not underlie experimental variation.

constitutional measures for unsaturation is the multiple bond count (*b**), which was expressed by [175] as

$$b^{*} = \sum\nolimits_{b} \left(\pi_{ij}^{*}\right) - B$$

where *B* represents all bonds in the molecule and π^* is the conventional bond order. [28]

4.2.1.3 Rotatable bonds count

Rotatable bonds count is the sum of bonds, which can freely rotate around themselves, giving indications on molecular flexibility. A rotatable bond is typically a single bond between two non-terminal heavy atoms.[29, 30] Moreover, they should not be part of a ring structure. However, potentially rotatable bonds like hydroxyl or methyl groups are often not included in calculations. Several QSAR studies imply that molecular flexibility is an important feature to describe interactions with biological targets.[177-179]

4.2.1.4 Molecular weight

Molecular weight (MW) is probably the simplest measure of molecular size. In contrast to simple atom or bond count descriptors the feature holds information on atom types. It is easily calculated by summing up the individual atomic weights of the compound, which is expressed as

$$\mathrm{MW} = \sum_{i=1}^{A} m_i$$

where *i* runs over all atoms (*A*) in the molecule and *m* is the atomic mass. The formula is adapted from [175]. Molecular weight is despite its simple calculation a fundamental parameter to describe biological and pharmacological behavior of compounds, as size plays a crucial role in permeation capacity and passive diffusion. Several rules of thumb use molecular weight to determine drug absorption and permeation (Section 3.2.1.1 and 3.2.1.2).

[28] Note that for saturated molecules, b* = 0.

[29] Non-terminal can also designate, in this context, a heteroatom connected to hydrogen.

[30] This bond should not be a triple bond, unless it is connected to another atom.

4.2.1.5 Partition coefficient (LogP)

The partition coefficient P measures the distribution of a compound between a hydrophobic and an aqueous phase.[31] It is probably the oldest and most widely used measure for lipophilicity. For easier handling, in place of P, its decadic logarithm LogP is commonly used.

Lipophilicity substantially influences a compound's distribution within the body. Hydrophobic compounds will eagerly permeate through lipid bilayers and enrich in lipidic environments (e.g., CNS), while their hydrophilic counterparts will distribute in aqueous compartments (e.g., blood serum). Several computational methods were proposed to calculate LogP. The most widely used ones are atom-centered (aLogP, xLogP) and fragment centered approaches (cLogP).

Moriguchi proposed a very generalized method to assess partition coefficients computationally (xLogP).[180] He proposed a regression analysis, where 13 structural elements of each molecule are determined and weighed in an equation. Ghose and co-workers introduced an atom-centered method, which considers the lipophilic contribution of each atom in dependence of immediate atomic neighborhood (aLogP).[181, 182] Both methods can be applied to a wide spectrum of molecules regardless of their complexity but in certain cases at the expense of accuracy. However, it can be helpful to get a rough estimate at very low computational expense.

The hydrophobic fragmental constants proposed by Leo and Hansch are probably the most accurate way of determining LogP values.[183] Non-overlapping fragments are generated by a simple set of rules, and their fundamental hydrophobic constants are determined. For simple compounds only containing one functional group this method is very accurate. For more complex structures, containing more than one functional group, correction factors were derived to improve LogP prediction.[32] Its computational implementation by Chou and Jurs became known as the calculated LogP (cLogP).[184]

4.2.2 Electronic Descriptors

Distribution and amount of a molecule's electricity is fundamental for its reactivity and behavior in chemical and biological systems.

4.2.2.1 Hydrogen bonding descriptors

Hydrogen bonding can be described as a dipol-dipol interaction between a hydrogen atom and electronegative atoms, which are usually constituted by oxygen, nitrogen or fluorine. Although hydrogen bonds are not as strong as covalent binding, they are still stronger than van der Waals interactions.

[31] The lipid phase has changed over time from olive oil to octanol and n-alkanes.

[32] This extensive list of fragments and correction factors holds information on proximity effects, hydrogen bonding, branching and many more.

However, the strength and distance of the bond depends on the kind of electronegative atom involved.[33]

Hydrogen bonds occur when strong positive molecular charge attracts a lone pair of electrons on a heteroatom. Generally, we designate a heteroatom with covalently bound hydrogen as a hydrogen bond donor. A heteroatom with a lone pair of electrons is termed hydrogen bond acceptor. However, a hydrogen bond donor can also accept hydrogen bonds and vice versa. Carbon can principally also participate in hydrogen bonding, if it is bound to an electronegative atom which decentralizes the electron cloud, leaving the molecule with a positive partial charge.

In a pharmacological context, a compound with strong hydrogen bonding capacity shows reduced permeation capacity. In order to permeate lipidic membranes hydrogen bonds have to be broken from the watery phase of e.g., blood serum, which is an energy-consuming step. This fact found reflection in several rules of thumb. For example, Lipinski’s Rule of Five generally associates high hydrogen donor and acceptor counts with bad "drug-ability" and brain permeation. Notably, the number of hydrogen bond donors is considered to be more detrimental for brain penetration than the number of hydrogen bond acceptors.[186, 187]

4.2.2.2 Charged partial surface area

Stanton and Jurs introduced charged partial surface area (cPSA), which describes the distribution of charge on the molecular surface. In this way, the descriptors consider features responsible for polar interaction between molecules. The molecular surface was defined as the overlap of the atomic van der Waals radii, which is traced by a sphere, representing a solvent molecule.[165] The molecular electron distribution is then projected on this accessible surface area.[188] Stanton and Jurs derived 25 descriptors that combined the solvent accessible surface with partial atomic charge. Table 5 summarizes the descriptors as they are implemented in the Chemical Development Kit (CDK) (Section 4.6.3).[189]

[33] The bond strength varies between 155 kJ/mol of fluorine bound hydrogen to fluorine and 8kJ/mol of nitrogen bound hydrogen to oxygen.[185]

Descriptor	Summary
pPSA1, pNSA1	Partial positive and negative surface area
pPSA2, pNSA2	Partial positive and negative surface area multiplied by the total positive charge on the molecule
pPSA3, pNSA3	Charge weighted partial positive and negative surface area
dPSA1	Difference of pPSA-1 and pNSA-1
dPSA2	Difference of fPSA-2 and pNSA-2
dPSA3	Difference of pPSA-3 and pNSA-3
fPSA1, fPSA2, fPSA3	pPSA1, pPSA2, pPSA3 / total molecular surface area
fNSA1, fNSA2, fNSA3	pNSA1, pNSA2, pNSA3 /total molecular surface area
wPSA1, wPSA2, wPSA3	pPSA1, pPSA2, pPSA3 multiplied by total surface area, divided by 1000
wNSA1, wNSA2, wNSA3	pNSA1, pNSA2, pNSA3 multiplied by total surface area, divided by 1000
rPCG, rNCG	Relative positive and negative charge
rPCS, rNCS	Relative positive and negative charged surface area
tHSA	Sum of solvent surface area atoms with partial charge less than 0.2
rHSA	tHSA / total molecular surface area

Table 5 - Listing of cPSA descriptors as they are implemented in the Chemical Development Kit (CDK).

In 2000, Ertl and co-workers proposed a fragment-based method to assess PSA, called TPSA. Single polar fragments are summed up to calculate surface contribution.[190] There exist two options for TPSA computation. The first variant considers only strongly polarizing fragments, which contain nitrogen and oxygen groups (TPSA[NO]). The second option considers additionally weak polarizing fragments like sulfur and phosphorus (TPSA[tot]).[175]

The topological method (TPSA) may be superior to conventional cPSA calculations regarding computation time,[34] and in addition, one does not require a 3D molecular geometry as it relies on a set of predefined polar features. However, a discrimination of positive and negative charge is not performed by TPSA, which reduces its information content. Moreover, its applicability for higher

[34] Ertl stated a two to three order of magnitude decrease in computation time.

molecular structures such as antibodies or proteins remains questionable, as in the initial study exclusively small molecules and drug-like structures were considered.[35]

4.2.3 Topological descriptors

Topological or connectivity descriptors are derived from the molecular graph. Topological indices are typically graph invariants, which means that they are not conformation or representation sensitive. Weighting schemes for edges or vertices can add additional information.

4.2.3.1 Adjacency matrix

The adjacency matrix is a fundamental graph theoretical matrix. It considers the immediate neighbor hood of single atoms in a molecule, i.e., whether two vertices are connected by an edge or not. The matrix entries equal one (1), if two vertices are connected and zero (0) if they are not.

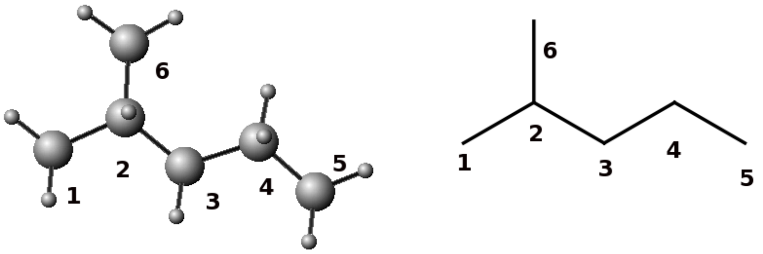


Figure 7 - Atom numbering of 2-Methylpenthane and corresponding molecular graph is shown.

Although the symmetric adjacency matrix (Table 6) does not account for multiple bonds, this information can be introduced by adding weighting schemes. From the adjacency matrix, we can derive the vertex degree (δ) by adding up each row.[175]

[35] Accessibility of polar substructures could be reduced due to folding effects.

Atom	1	2	3	4	5	6	δ
1	0	1	0	0	0	0	1
2	1	0	1	0	0	1	3
3	0	1	0	1	0	0	2
4	0	0	1	0	1	0	2
5	0	0	0	1	0	0	1
6	0	1	0	0	0	0	1

Table 6 - Adjacency matrix with corresponding vertex degree (δ) is given for the molecule shown in Figure 7.

From the adjacency matrix, path counts can be derived, where the path from the vertex *i* is counted to any other vertex in the graph. The path order is defined as the length of the path.

4.2.3.2 Weighted matrices

The principle of an augmented adjacency matrix was proposed by Randic.[191-193] He replaced zero values from the symmetry axis of the adjacency matrix by characteristic values for atom types in the molecule (e.g., physicochemical properties). Its vertex degree is analogously called augmented vertex degree.

The Burden matrix (Table 7) and its eigenvalues were proposed by Burden in 1989.[194] In analogy to the augmented adjacency matrix, the Burden matrix replaces the diagonal zeros by atomic numbers of the corresponding atoms. The edges are weighted by their corresponding bond order, taking also aromaticity into account.

Atom	1	2	3	4	5	6	δ
1	6	0.11	0	0	0	0	6.11
2	0.11	6	0.1	0	0	0.11	6.23
3	0	0.1	6	0.1	0	0	6.2
4	0	0	0.1	6	0.11	0	6.21
5	0	0	0	0.11	6	0.01	6.12
6	0	0.11	0	0	0.01	6	6.22

Table 7 - The Burden matrix is given for the molecule depicted in Figure 7. Here, diagonal elements are the atomic numbers (i.e., carbon). If atoms are connected, conventional bond order is put into the matrix (i.e., 0.1 for single bonds, 0.2 for double bond, 0.3 for triple bonds and 0.15 for aromatic bonds). Terminal bonds are augmented by 0.01.

A popular extension of the Burden matrix are the eigenvalue-based BCUTS descriptors, where diagonal elements are replaced by varying weights.[195]

4.2.3.3 Kappa shape indices

Kier proposed in 1985 and 1986 the kappa shape indices κ, which relate the hydrogen depleted chemical graph (P_i) to a minimally (P_{min}) and a maximally (P_{max}) connected reference graphs in a way that their relationship holds as follows[196]:

$$P_{\min} \leq P_i \leq P_{\max}$$

By putting the chemical graph in relation to different reference graphs, Kier proposed three kappa indices, which give information on different aspects of molecular shape. For the κ1 index, the minimal graph was defined as the linear graph, while the maximum graph is the complete graph, where every vertex is connected to each other. The information rising form κ1 is related to numbers of cycles in a molecule (Figure 8).

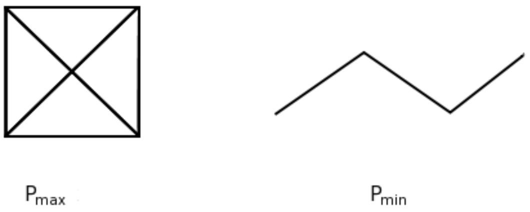


Figure 8 - For κ1, the maximal and minimal graph (P_{max} and P_{min}) are the complete and the linear graph, respectively.

The κ2 index measures spatial distribution of atoms in a molecule. Reference graph extremes are the linear and the star graph (Figure 9).

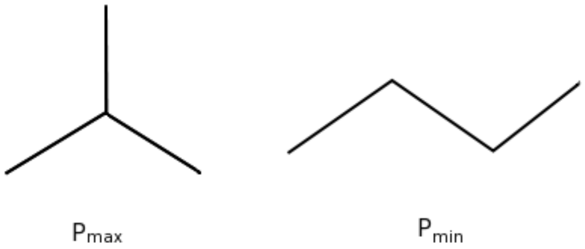


Figure 9 - For κ2, the maximal graph (P_{max}) is the star graph, while the minimal one (P_{min}) is the linear graph.

The κ3 index encodes information on centrality of molecular branching, as its values increase when molecules are not branching or only branching at their extremities. Upper limit is the twin star graph while the under limit is the linear graph (Figure 10).

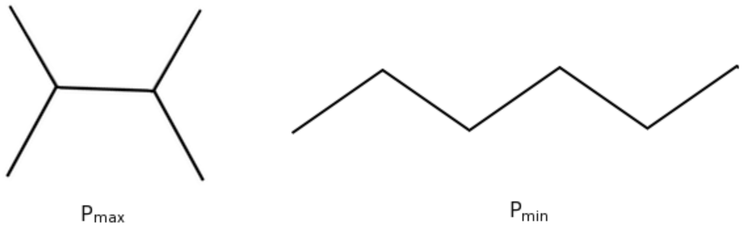


Figure 10 - For κ3, the maximal and minimal graph (P_{max} and P_{min}) are the twin star and the linear graph, respectively.

4.2.4 Geometrical Descriptors

Although geometrical descriptors differ in definition, they all deal with actual molecular spatial distribution and shape. Usually, 3D-coordinates are derived from computational force fields or from christallographic data. An example is the length over breadth descriptor, where maximum and minimum ratio of molecular length and breadth are considered. Other typical measures are gravitational indices, or the principal moments of inertia.

4.2.4.1 Gravitational Index

Gravitational indices give information on intramolecular mass distribution. In other words, they describe molecular density and cohesion. Molecules can be considered either with or without hydrogen atoms. Gravitational Indices four to six consider all atom pairs, regardless of whether they are bonded or not. Wessel and co-workers also proposed the use of the square and cubic root of the descriptors.[197]

4.2.4.2 Principal moments of inertia

Another way of quantifying mass distribution is the consideration of molecular rotational dynamics by the principal moments of inertia. According to Todeschini and co-workers the moment of inertia for any of the three principal axes X, Y, and Z is defined as

$$I = \sum_{i=1}^{A} m_i \times r_i^2$$

A is the atom number of a molecule, *i* stands for the *i*-th atom in that molecule, while *m* is the atomic mass and *r* is the perpendicular distance to the considered axis.[175] Moment of inertia, calculated along the three principal axes as well as their ratios are used for modeling.

4.2.4.3 Radius of Gyration

Closely related to moments of inertia is the radius of gyration (R_G). The descriptor assesses the molecular compactness by relating atomic distance from the center of molecular mass to molecular weight. This can be formulated as follows, adapted from [198]

$$R_G = \sqrt{\frac{\sum_{i=1}^{A} m_i \times r_i^2}{MW}}$$

Analogously to the formula of the moment of inertia, *A* is the atom number of a molecule, *i* stands for the *i*-th atom in the molecule, while *m* is the atomic mass and *r* is the perpendicular distance to the considered axis. Molecular weight is abbreviated as *MW*.

4.2.4.4 Petitjean Shape Indices

Petitjean proposed in 1992 the shape coefficient (*I*), which measures molecular anisotropy based on the graph theoretical approach. He used minimal (generalized radius [*R*]) and maximal (generalized diameter [*D*]) paths in the molecule and defined the following relationship:[199]

$$I = \frac{(D-R)}{R}$$

He suggested that this index would correlate graph theoretical and geometrical shapes. Bath proposed the geometrical shape index by extending Petitjean’s principle. He applied the geometrical matrix instead of a graph theoretical one.[200] However, he relativized Petitjean's claims by his observation that there is only a low degree of correlation between these two measures.

4.3 Machine learning paradigms

4.3.1 Decision tree induction (DTI)

Human learning is characterized by splitting problems into smaller sub-problems, in order to ease classification. This principle is mimicked by decision tree induction (DTI). The paradigm is efficient and powerful in solving even non-linearly separable problems. Moreover, the trees branches can be read as single rules, which eases practical implementation and their use to predict future instances. A tree grows by splitting data on its attributes (i.e., splitting criteria) in smaller subsets (Figure 11). This process is then recursively repeated on the subsets, until a certain degree of purity is achieved. This process is termed recursive partitioning. Tree growth is terminated, if either the leaves contain only one class (e.g., have reached perfect purity) or further splitting does not improve purity.

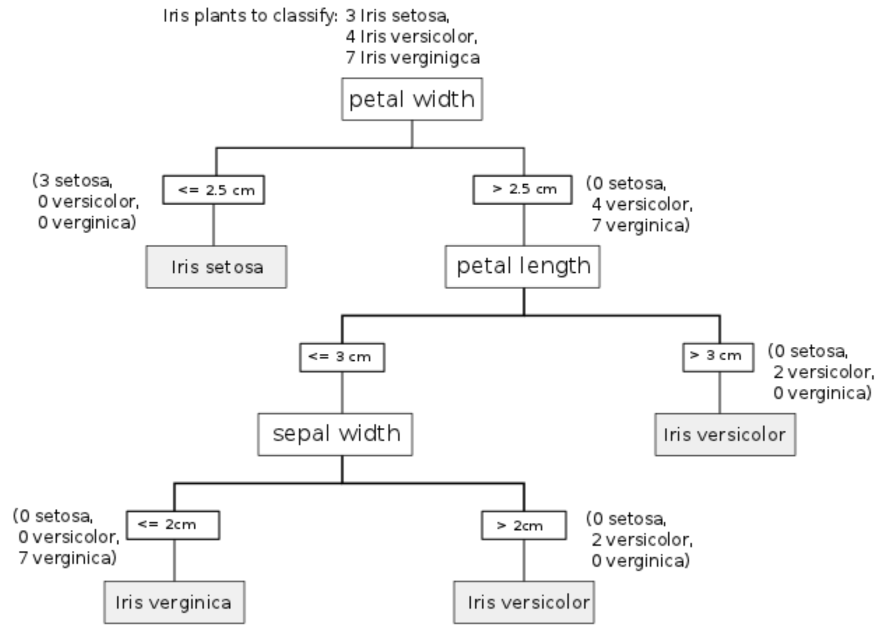

Iris plants to classify: 3 Iris setosa, 4 Iris versicolor, 7 Iris verginigca
petal width
(3 setosa, 0 versicolor, 0 verginica)
<= 2.5 cm
> 2.5 cm
(0 setosa, 4 versicolor, 7 verginica)
Iris setosa
petal length
<= 3 cm
> 3 cm
(0 setosa, 2 versicolor, 0 verginica)
sepal width
Iris versicolor
(0 setosa, 0 versicolor, 7 verginica)
<= 2cm
> 2cm
(0 setosa, 2 versicolor, 0 verginica)
Iris verginica
Iris versicolor

Figure 11 - A dummy decision tree of the famous Fisher's Iris data set is shown.[201]

Concerns were raised, that DTI is highly vulnerable and unstable when no stratification is used in generating test and training sets.[202] Perturbing the training set would then cause significant changes in the predictor. To avoid such shortcomings, one could typically use y-scrambling, where endpoints are randomly perturbed. The model trained on such data can then be used to uncover randomly correlating features.

4.3.1.1 Pruning

The aim of pruning is to simplify the final tree and improve its predictivity by reducing overfitting and noise. Pruning generally replaces a node or a whole subtree with a leaf,[36] sometimes at the expense of accuracy on the training set but for the sake of avoiding overfitting on unseen instances. The decision whether a node shall be replaced or not, could be made by comparing the error on the hold out data of the pruned and unpruned tree. If the error becomes smaller, the original tree will be pruned. Pruning can be applied during tree growth (forward pruning), which is advantageous as it would avoid time-consuming subtrees growth for futile branches. However, post pruning considers the tree after its complete building. It offers the possibility to overcome situations, where single splitting attributes have less discriminatory power than the consecutive combination of them. In this case, the branch would have been prematurely terminated by forward pruning. Most DTI algorithms apply backward pruning.

[36] In subtree raising, nodes are replaced with nodes below them.

4.3.1.2 Classification and regression tree (CART)

Leo Breimann introduced in 1984 the classification and regression tree algorithm (CART). It produces binary trees, which can handle categorical and ordinal and continuous data. Depending on the data trained on, CART builds regression or classification trees. CART uses backward pruning.[203] For splitting criterion selection, CART maximizes Gini impurity.[204, 205] The Gini impurity (also called Gini coefficient) is a measure for curve deviation from chance line. It gives indications on the distance of a curve (e.g., ROC curve or a cumulative curve) to the chance line, which allows drawing conclusions on its slope or it's information content (Figure 12).

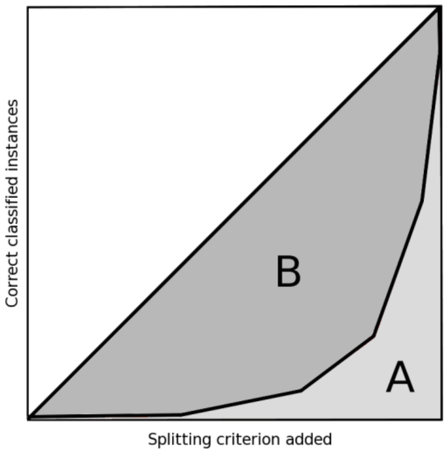


Figure 12 - The curve illustrates the improvement of classification by splitting criteria added. The diagonal is the chance line or the line of no discrimination. A is the area under the curve, while A+B indicates the area under the chance line which equals 0.5.

The diagonal line of the ROC area indicates the chance line or perfect equality in cumulative curves. The area under this curve equals 0.5. The Gini impurity can accordingly be formalized as

$$\text{Gini impurity} = \frac{|A - 0.5|}{0.5}$$

where *A* stands for the area under the curve. This formulation is independent of the curve's deviation (i.e., concave or convex). The Gini impurity is scaled from 0 to 1, where 0 stands for total equality and the value 1 stands for perfect inequality. CART applies the coefficient to decide whether a criterion is worth splitting on. A good splitting criterion with high discriminatory power has preferably a value near 1. This indicates that it has a very unequal distribution in the dataset, i.e., it will have a high discriminatory power.

4.3.1.3 Chi-squared automatic interaction detector (CHAID)

CHAID is the oldest DTI paradigm, proposed in 1964 by Sonquist.[206] Typically, CHAID uses forward pruning, in contrast to most of the other DTI algorithms. Attributes for splitting are chosen by the chi-square test. The chi square test rates which child node adds the most information to the tree. The null hypothesis states that there is no difference in information between child and parent node. With an increasing chi-square, the information a child node adds to the tree diverges from the null hypothesis, hence a real information gain exists. The node with highest chi-square, i.e., the feature that adds the maximum information compared to the parent is selected to split on. In contrast to CART, CHAID handles exclusively categorical data.

4.3.1.4 Random Forests

In 2001, Breimann and Cuttler introduced the principle of Random Forests (RF).[207] As the name implies, Random forests consist of multiple decision trees. Principally, any DTI paradigm can be use to create a random forest. Each tree is grown to maximal depth with a randomly composed subset of all available features and is not pruned.[208] The final prediction is yielded by a majority vote of all trees on the final classifier. Random forests were praised as outperforming many other machine learning paradigms and efficiently handling enormous data and feature sets. However, they are suspected to be susceptible to overfitting and noise.[209][37]

4.3.2 Artificial neural networks (ANN)

The ANN paradigm is an abstraction of a biological network of neurons. Instances are represented as vectors containing their features.[210] Each feature is passed to one of the input neurons to which a weight is assigned. Based on these weights, input is passed to the output layer over a number of interspersed optional hidden layers. The output layer combines these signals to produce a result. Initially, weights are set to random values. As the network is repeatedly presented with training instances, these weights are adjusted so that the total output of the network approximates the observed endpoint values associated with the instances.

4.3.3 Support vector machines (SVM)

Support vector machines (SVM) were introduced by Cortes and Vapnik.[211] A major advantages of SVM are their low computational expanse, as they do not search for separating hyperplanes by considering all instances but only those data points which confine borders of classes. Moreover, the paradigm exhibits an extraordinary robustness concerning classification of noisy data as the separating "solution plane" is spanned with maximal distance to the class borders of the training set, thus allowing also correct classification of instances lying even nearer to the decision plane than instances from the original training set did. The output of the SVM is basically a plane equation, which is solved for new instances as either >1 or <-1, while for instances exactly on the plane it is 0.

[37] Segal argued that Breiman used datasets for testing RF which could hardly be overfitted.

As SVM is in principle a linear paradigm, it could be considered useless in solving non-linear problems. However, in theory for any non-linear problem, a linear solution can be found by sufficient (some times up to infinite) up-transformation of the original feature space. It is evident, that such models would become difficult to computationally handle as well as they would increasingly suffer from an overfitting bias. To avoid unreasonably high dimensionalities of the feature space, kernel tricks are usually applied.[212] Kernels implicitly transform the feature space into an inner product space searching for meaningful linear solutions, which can then be computed without an explicit transformation of the original space. Common kernel functions include the polynomial kernel

$$K(x_i,x_j) = (\gamma x_i x_j + \mathrm{const})^d$$

and the radial basis function (rbf) kernel.

$$K(x_i,x_j) = \exp(-\gamma \left\| x_i - x_j \right\|^2)$$

In SVM with rbf kernels, there are two learning meta-parameters that greatly influence performance (Cost [c] and gamma [γ]). Determining these parameters is an optimization task.

4.3.4 Naive Bayes

Thomas Bayes' (1702-1761) theorem was posthumously published and was to revolutionize the doctrine of probability.[38] He stated that one could deduce the conditional probability Y given X (Y|X)[39], when we know the unconditional (prior) probabilities of X and Y and of their conjunction, the conditional probability X given Y (X|Y). Then the following statement holds true

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Bayes' theories had a revival in the 1950's and proved especially useful in conjunction with Markov chain methods (Bayesian networks).[214, 215] Additionally, Naive Bayes are very robust against missing values: the probability ratios are based on the actual number of occurrence and not on the instance number. On the other hand, one must take care that none of the probabilities equals zero. This could be the case if a particular attribute or condition does not occur in conjunction with the other one. Hence, this particular fraction would equal zero and due to multiplications, the final estimate would have a zero occurrence. To avoid such shortcomings one should introduce an a priori probability to every case, also known as Laplace estimator.

[38] It is thought that Thomas Bayes, an English mathematician and Presbyterian minister, did not publish his observations during his lifetime as he calculated a value of less than 1 for the probability of god's existence. In fact, Stephen D. Unwin used his theorem and calculated a probability of 67% that god exists.[213]

[39] (Y|X) is also called the posterior probability.

Naive Bayes presumes independence of prior probabilities. However, this precondition is rarely fulfilled by real world data. Therefore one could argue that such a simplified approach would produce rather over-optimistic classifiers. However, Naive Bayes predicted very strong on realistic data and frequently outperformed rather sophisticated machine learning paradigms (Section 5.1).[216, 217] Zhang discussed reasons and conditions under which this effectiveness cannot be accounted to overfitting. He stated, that if variable dependence is equally distributed between classes or is canceled out, Naive Bayes would produce reasonable results.[218] Another explanation was proposed by Domingos, who argued that Naive Bayes' performance is owed to the so-called zero-one loss function.[219] This function defines the error as the number of incorrectly allocated class labels to classified instances.[220] This means that Naive Bayes are still able to assign the correct class to an instance, although the exact probability estimate might be poor.[219] In accordance to Ockham's Razor, simple solutions should be preferred to complicated ones and Naive Bayes is therefore an elegant technique worth considering. In fact, its intriguing clearness and its robustness to missing values makes it a good choice for machine learning.

4.3.5 K-nearest neighbor

The K-nearest neighbor (KNN) paradigm is a lazy learning paradigm. This means the classifier does not produce a model in advance but compares new instances at runtime with its known instance space, which is spanned by its instance database. New, unseen instances are assigned to the class of its immediate neighborhood.[221] The concept of neighborhood can be measured in various ways of which two are illustrated in Table 8.

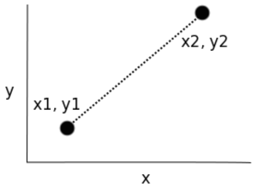
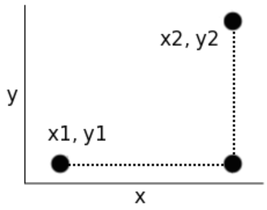
Distance measures	Illustration	Equation
Euclidian distance		$\sqrt{(x1-x2)^2+(y1-y2)^2}$
Manhattan/city block distance[40]		$(x1-x2)+(y1-y2)$

Table 8 - Euclidian distance and Manhattan or city block distance are illustrated.

[40] Manhattan distance approaches Euclidian distance with increasing resolution.

4.3.6 LAZAR

The LAZAR engine [222] is a lazy learning fragment-based predictor. Relevant fragments are determined by finding all linear fragments in the training dataset (without size limits) and removing those that are statistically insignificant ($p<0.95$) in the chi-square test. Remaining relevant fragments are used to determine activity-specific similarities of all compounds in the training set with a weighted Tanimoto index. LAZAR classifies unknown compounds with a modified k-nearest-neighbor (KNN) algorithm.

4.3.7 Notes on lazy learning

In contrast to methods which produce a model during the training period (i.e., eager learning methods), lazy learning paradigms compare unseen instances at runtime to their knowledgebase. Therefore, lazy classifiers do not use any information compression. This is an advantage, when new instances should be incorporated into the classifier: the paradigm can be easily extended by simply adding these new instances to the database. However, for the same reasons, lazy learning classifiers have a reduced portability compared to models produced by eager learning paradigms.

4.4 Feature selection and optimization tasks

Without reducing the abundance of descriptors we are able to generate, one runs the risk of detecting meaningless correlations with highly accurate predictions, i.e., overfitting models. The obvious solution is to propose a hypothesis of the relationships involved. This should be in fact the starting point of every statistical analysis of data. Ideally, we base feature selection on our mechanistic knowledge of the process which should be modeled. Once we have reduced the feature space, we could still end up with too many descriptors compared to the size of the dataset. Generally, we should adapt the feature number to the instance set size. For this purpose, a feature reduction algorithm could be helpful. These paradigms are designed to pick out the variables with low intercorrelation and strong explanatory power.

Best first feature selection (BFS) searches the feature space for the best combination of samples, continuously expanding the feature set (or reducing it, depending on the direction of the search). Forward selection starts with comparing all features in isolation and selecting the best performing one, according to BFS's heuristic function. To this selected feature, one of the remaining features is combined and again tested, incrementally expanding the feature set. When no more improvement can be obtained by adding new features the search is terminated. Backward selection starts with the whole feature set, reducing it by one feature and testing whether the new set performs better. If this is the case, the set is decremented and tested again. The paradigm terminates the search if the result cannot be improved. Accordingly, backward search will generally end up with bigger features sets than forward selection.[223]

Feature selection is essentially an optimization problem, where we do aim to find a solution, but will not mandatorily end up with the best one. At first glance, this might be disappointing; however, although available computing power is rapidly growing, there are many problems which are inherently non-solvable in reasonable time. For these so-called nondeterministic polynomial problems (NP), the best solution cannot be found in on polynomial time. Although we cannot find the best solution, we still are capable of finding a reasonable solution by a heuristic optimization. Heuristic procedures are often more reliable and robust than searching for the best solution (Figure 13).[41]

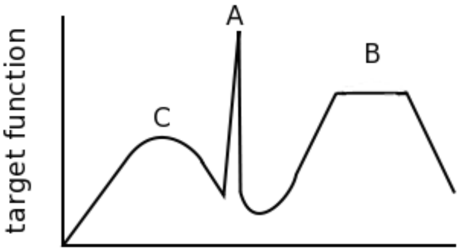


Figure 13 - A target function is shown. Solutions for the local maxima (C and B) are easier found and most likely more robust than the solution on the global maximum (A). Minor deviations of the target function from the global maximum can lead to a substantially decreased performance. If a solution is found on a plateau (B), the function will perform robustly even if deviations in the target function do occur.

A typical example of an optimization problem which is NP-hard is the "traveling salesman"-problem: given a list of cities and their corresponding pairwise distance, the shortest tour has to be found, visiting every city only once (Figure 14).

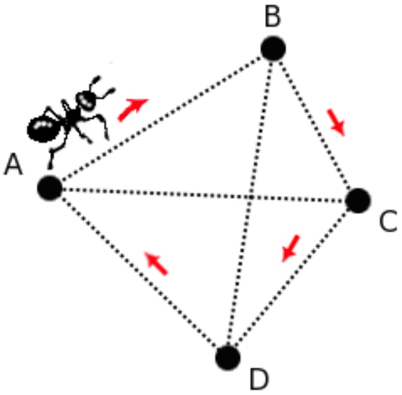


Figure 14 - An abstraction of the traveling salesman problem is shown. The agent (here an ant) has to visit every point only once, but using the shortest route.

[41] E.g., in engineering, robust solutions are preferred to isolated global maxima as in real life applications deviations in parameters of the target function are often seen.

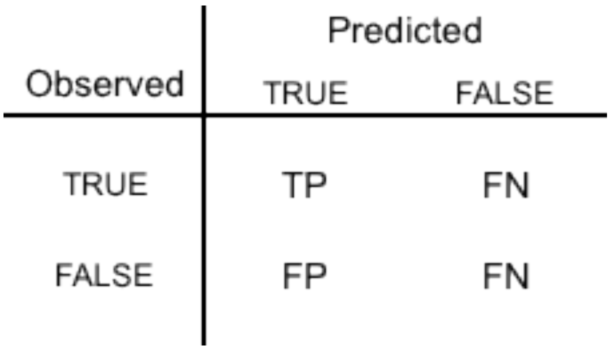
In recent years, the field of natural computing has produced intriguing heuristics for such optimization problems. Ant Colony Optimization (ACO), was introduced in the 1990s, where real world ants foraging behavior is simulated.[224] When real-world ants find a food source, they will return in a more or less direct way back to their colony, marking the path with a pheromone track, which should guide other ants to the same food source. Pheromones are subjected to evaporation, which will eventually lead to a preference for shorter paths as more and more ants will use the one, which exhibits the most pheromone. In this way, the colony exhibits a tendency to converge. Once convergence is reached, ants will be unlikely to explore other paths. This holds for virtual as well as for real-world ants: in the so-called "double bridge" experiment, an obstacle hinders direct access to a food source. The ants will nevertheless find a path around it and the path will then be reinforced. However, when the obstacle is removed, the ants will still follow the prior, and now suboptimal, path.[42]

In ACO, ants are abstracted agents scurrying about a graph at random until finding a solution (a food source). Other ants explore the graph and weigh their choices of route by previously deposited pheromones. The optimization has been successfully applied to a fragment based feature selection task (Section 5.5).

4.5 Quality measures

4.5.1 Confusion matrix and derived metrics

Results from classification can be represented as a contingency or confusion matrix (Table 9). Each row represents instances in a predicted class, while each column represents instances in the actual class. In this way, exact numbers of truly classified positives (TP), truly classified negatives (TN), falsely classified positives (FP) and falsely classified negatives (FN) are presented in tabular form.



	Predicted	
Observed	TRUE	FALSE
TRUE	TP	FN
FALSE	FP	FN

Table 9 - The confusion matrix is depicted. In TP and FP indicate true and false positive instances, FP and FN true and false negative instances, respectively.

[42] The experiment was conducted using the argentinian ant species Iridomyrmex humilis [225] Linepithema humile, and Lasius niger.[226]

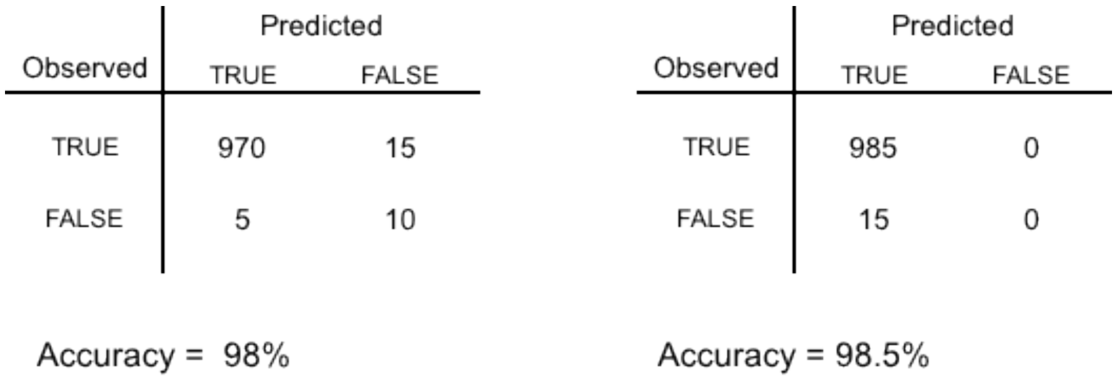
In presence of a confusion matrix one can easily compute quality measures of which we will discuss the most important.

Accuracy (Formula 1) is the measure of proximity of a value to its actual (true) value. It is a common measure to start performance analysis.

$$\text{Accuracy} = \frac{T_P + T_N}{N_0 + N_1}$$

Formula 1 – Accuracy, where TN and TP stand for true positive and negative instances. N_0 and N_1 stand for all positive and all negative instances.

However, the measure underlies the accuracy paradox, which states that a model with a given accuracy might have a higher predictive power than a model with higher accuracy. This problem arises as the measure does not take false classified instances into account. A model predicting a highly unbalanced dataset is considered below (Table 10). Assigning all instances to the majority class would produce a useless model, but with higher accuracy than in highly unbalanced datasets to predict.



	Predicted	
Observed	TRUE	FALSE
TRUE	970	15
FALSE	5	10

Accuracy = 98%

	Predicted	
Observed	TRUE	FALSE
TRUE	985	0
FALSE	15	0

Accuracy = 98.5%

Table 10 - The confusion matrix for a hypothetical classification is given to illustrate the accuracy paradox. Accuracy of both predictors is indicated. The right example assigns all instances to the majority class resulting in a useless model, but exhibiting a higher accuracy than the left predictor, which has actually discriminatory power.

Precision or positive predictivity is a measure for reproducibility and assesses the degree of variance in measurements performed repeatedly under the same conditions.

$$\text{Precision} = \frac{T_P}{T_P + F_P}$$

Formula 2 - Precision TP and FP stand for true and false positive classified instances.

It is the proportion of positive test results, which were correctly assigned. In a clinical context, it is a most important quality measure for diagnostic tests as it reflects the probability that a positive result reflects a tested condition. It is important to note that a method or a test can exhibit high accuracy but low precision and vice versa (Figure 15). However, it would be desirable to have both, high precision and high accuracy.

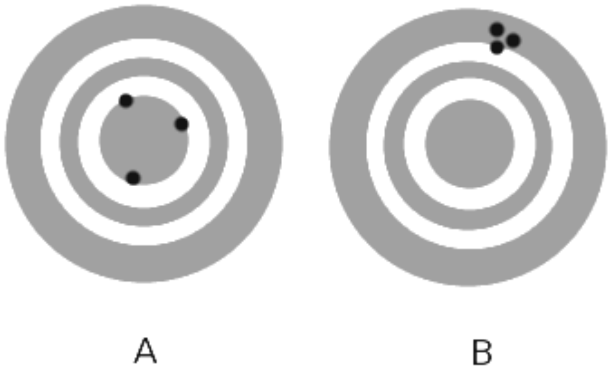
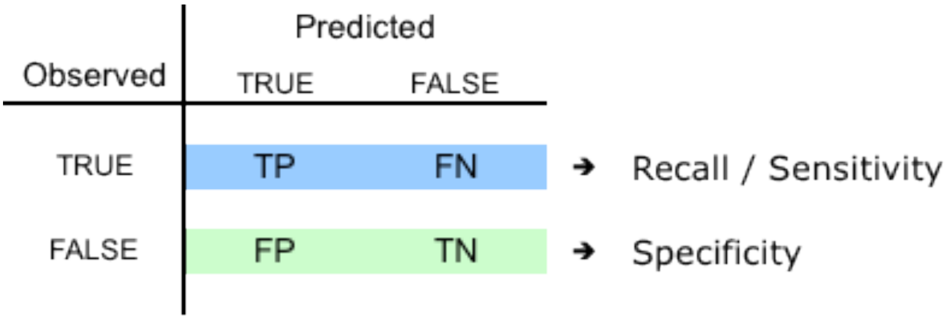


Figure 15 - Target A shows bullet holes with high accuracy and low precision. The holes in target B show a high precision, but low accuracy.

Recall or sensitivity measures the rate of positive instances classified positively, while specificity assesses the correctly classified negative instances. Although both performance estimates give an intuitive quality estimate, they should not be used independently of each other.



	Predicted			
Observed	TRUE	FALSE		
TRUE	TP	FN	→	Recall / Sensitivity
FALSE	FP	TN	→	Specificity

Table 11 - Confusion matrix is shown to illustrate the composition of the quality measures recall/sensitivity and specificity.

The formula for recall/sensitivity and specificity are indicated below:

$$\text{Recall} = \frac{T_P}{T_P + F_N}$$

$$\text{Specificity} = \frac{T_N}{T_N + F_P}$$

4.5.2 Corrected classification rate and Matthews correlation coefficient

It is trivial, that assignment of all instances to the majority class, (i.e., a perfect overfitting), would produce a recall/sensitivity of 100%, which overestimates the actual predictive power (i.e., accuracy paradox). For this reason, accuracy, recall/sensitivity, and precision are usually not considered in isolation. Especially in unbalanced datasets, where overfitting might easily occur, the use of measures that consider also falsely classified instances is an advantage. Usually, a combined measure such as the corrected classification rate (CCR) or Matthews correlation coefficient (MCC) is used.

$$\text{CCR} = \frac{1}{2}\left(\frac{T_N}{N_0} + \frac{T_P}{N_1}\right)$$

Formula 3 - CCR, where T_N and T_P refer to compounds classified as true negatives and true positive instances. N_0 are all negative and N_1 are all positive instances.

$$\text{MCC} = \frac{T_N T_P - F_N F_P}{\sqrt{(T_P + F_P)(T_P + F_N)(T_N + F_P)(T_N + F_N)}}$$

Formula 4 - In the MCC formula, additionally falsely negative (FN) and falsely positive (FP) classified molecules are considered.

4.5.3 Receiver operating characteristics (ROC)

Another approach to analyze classification performance is the use of receiver operating characteristics (ROC) curves.[43] ROC graphs depict relative tradeoffs between benefits (true positives, TP) and cost (false positives, FP) in a two-dimensional plot. Sensitivity is usually depicted on the ordinate and the reciproque of specificity (1-specificity) is represented on the abscissa. In other words, every correctly classified instance will increase steepness of the discontinuous curve, while incorrectly classified ones decrease it by stepping further on the abscissa. A perfect classifier would yield a point in the upper most left corner, while points on diagonal dividing the ROC space (also known as chance line, or "line of no-discrimination") represents a complete random guess. All points lying under the no-discrimination line indicate poor classifiers, but can simply be inverted to become predictive.

One of the favorable advantages of ROC curves is their insensitivity to skewed datasets as they consider only TP and FP. For model comparison, the area under the ROC curve (AUC) is usually used. Interestingly, the AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen negative instance. This is equivalent to the Wilcoxon test of ranks and is closely related to the Gini coefficient (Section 4.3.1.2), which is twice the area between the diagonal and the ROC curve. Moreover, ROC curves are often used for model optimization tasks. This can be done by determination of optimal thresholds or cutoff points. A frequently used measure for evaluating model effectiveness is the J-index, first introduced in the medical literature by Youden.[227]

$$J = \mathrm{Sensitivity} + \mathrm{Specificity} - 1$$

Youden's J is measured over all cut points on the ROC curve to find the maximum vertical distance from the curve to the chance line in the upper left corner, ranging between zero an one. The intuitive interpretation of Youden's J index is that its maximal value is the point on the curve farthest from chance in the upper left corner of the ROC space.

4.6 Validation

4.6.1 Holdout validation

There are several approaches to validate machine-learning models. Holdout validation is a commonly used method, where the data is split in a test and training set. One-third is usually held out for testing, while the other two thirds are used for training. Holdout validation has the distinct advantage of taking a short time to compute. However, the drawback of this procedure is the difficulty to predict whether a sample drawn in such way will be representative for the data or not. When a class is underrepresented

[43] Introduced in world war II for analysis of radar signals, they were employed in the 1950 in psychophysics to assess human detection of weak signals. From there, ROC found extensive use in the medical field for evaluations of diagnostic tests and medical decision-making. Recently they have been used increasingly in machine learning and data mining.

or even missing in the training set, the classifier will not be able to consider it appropriately. Moreover, when validated, the method will almost completely fail to classify, as the missing class will be overrepresented in the test set. This holds also for the opposite: when test sets consist exclusively of the majority class, validation is prone to be over optimistic. One might address this issue by refining the partitioning scheme and distributing variance of individual attributes evenly over training and test set, e.g., using stratified datasets. However, there is considerable information crossover between both sets, again increasing the risk of overfitting.

4.6.2 K-fold cross-validation

In k-fold cross-validation, a data set is randomly divided into k subsets.[228] Of these subsets, k-1 sets are recombined to make up a training set, with the resulting model tested against the remaining instances. This procedure is repeated k times until all instances have served both as training and test data, thereby making sure that no classes are left out. This procedure is basically k-fold repetition of holdout validation. It is evident that it makes much more efficient use of the data. Consequently, independence on dataset composition increases and variance in performance is reduced the higher k is selected. The most extreme example is leave-one-out (LOO) cross-validation, where all instances are used for training except for one, which is used for testing. However, k is usually set to 10.

4.6.3 Software used

Molecular descriptors were generated with the open-source cheminformatics package Chemical Development Kit (CDK, Version 1.2.3, 2009, http://sourceforge.net/projects/cdk).[189] For several descriptors, 3D structures had to be derived from SMILES representations by the Ghemical force field (http://www.uku.fi/~thassine/projects/ghemical/).[229] We used Weka [223] (Version 3.6; http://www.cs.waikato.ac.nz/~ml/Weka/) for RF, SVM, ANN, KNN, and Naive Bayes. We performed DTI with PASW Statistics version 18 for Windows (http://www.spss.com/statistics/) and linear correlation analysis with R (http://www.r-project.org/). Chemical structure diagrams were created using ChemAxon MarvinSketch (Version 5.2.5; http://www.chemaxon.com/). LAZAR is available at http://www.in-silico.de/. Tanimoto coefficient calculations and grid screening for SVM meta-parameters were done with in-house software.

5 Projects

5.1 Combinatorial QSAR Modeling of human Intestinal Absorption

Claudia Suenderhauf [1], Felix Hammann [1], Andreas Maunz [2], Christoph Helma [2,3], and Jörg Huwyler [1]

[1] Pharmaceutical Technology
Department of Pharmaceutical Sciences, University of Basel, Klingelbergstrasse 50,
CH-4056 Basel, Switzerland

[2] Freiburger Zentrum f. Datenanalyse u. Modellbildung (FDM), University Freiburg,
Hermann Herder, Str. 3a, D-70104 Freiburg, Germany

[3] in silico toxicology, Altkircherstr. 3a, CH-4054 Basel, Switzerland

Molecular Pharmaceutics 2010 Dec 29.

Abstract

Intestinal drug absorption in humans is a central topic in drug discovery. In this study, we use a broad selection of machine learning and statistical methods for the classification and numerical prediction of this key endpoint. Our dataset is based on a selection of 458 small drug-like compounds with FDA approval.

Using easily available tools, we calculated one- to three-dimensional physicochemical descriptors and used various methods of feature selection (best-first backward selection, correlation analysis, and decision tree analysis). We then used decision tree induction (DTI), fragment-based lazy-learning (LAZAR), support vector machine classification, multilayer perceptrons, random forests, k-nearest neighbor and Naive Bayes analysis to model absorption ratios and binary classification (well-absorbed and poorly absorbed compounds).

Best performance for classification was seen with DTI using the chi-squared analysis interaction detector (CHAID) algorithm yielding corrected classification rate of 88%, (Matthews correlation coefficient of 75%). In numeric predictions, the multilayer perceptron performed best achieving root mean squared error of 25.823 and a correlation coefficient of 0.6. In line with current understanding is the importance of descriptors such as lipophilic partition coefficients (LogP) and hydrogen bonding. However, we are able to highlight the utility of gravitational indices and moments of inertia, reflecting the role of structural symmetry in oral absorption.

Our models are based on s diverse dataset of marketed drugs representing a broad chemical space. These models therefore contribute substantially to the molecular understanding of human intestinal drug absorption and qualify for a generalized use in drug discovery and lead optimization.

Introduction

See Section 3.2.1.1.

Materials and Methods

Dataset

The dataset used for the present study is based on a list of FDA approved small molecule drugs (n = 458) for which experimental data were available and sufficiently documented.[230] Omissions were due to missing information. Intestinal absorption (%Abs) is defined as the percentage of the dose absorbed from the gastrointestinal tract following oral administration. This is not necessarily the same as the amount of drug reaching systemic circulation, which is also affected by pre-systemic metabolism (e.g., hepatic first-pass effect). The arithmetical mean was used wherever an absorption range was given. We also did not omit compounds that are known substrates of efflux transporters such as P-glycoprotein (e.g., digoxin) because insufficient information on specific absorption and excretion pathways was supplied in the original data source.

As classification algorithms require nominal class labels as end points, we re-coded the numerical absorption ratios into three different ordinal classes (“TRUE”, “UNKNOWN”, “FALSE”, see Table 12). Thresholds were determined so as to produce sufficiently large number of instances for each class.

Class label	%ABS	Number of Instances
true	a ≥ 80%	303
unknown	30% < a < 80%	82
false	a ≤ 30%	73

Table 12 - Ordinal classes and corresponding absorption values (%Abs).

To achieve a better separability, members of the class "UNKNOWN" were exempted from classification learning. This class, corresponding to moderately absorbed compounds, was clearly underrepresented in the data source and hence was not deemed suitable for modeling. For numerical predictions, the entire dataset was used. Pre-study analysis (data not shown) indicated that data transformation, i.e. linearization, does not improve model performance.

Lastly, the data source provided generic drug names but no structural information. We therefore retrieved the corresponding structures from the National Library of Health database PubChem (http://pubchem.ncbi.nlm.nih.gov/). For salts, the counterion was removed prior to further processing.

Descriptors

See Section 4.2.

An overview is given in Table 13. For a full list and explanations see Section 4.2.

Class	Type
Charge Analysis	Hydrogen bonding capacity
	Charged partial surface descriptors
	Partitioning coefficients
	Molecular polarizability
Constitutional	Counts of atoms, rings, and bonds
	Length over breadth descriptors
	Gravitational indices
	Moment of inertia
	Molecular weight
Topological	Eccentric connectivity index
	Weighted Burden matrix
	Kier-Hall kappa shape indices
	Petitjean number and index
	Wiener path and polarity numbers
	Zagreb Index

Table 13 - Overview of descriptors (n = 80) used in this study. A detailed listing of all features is given in the supporting information.

The structural information retrieved from PubChem was two-dimensional. For certain descriptors, however, three-dimensional structures are required. We extrapolated these using OpenBabel (Version 2.2.3, http://www.openbabel.org/) to perform a search of lowest energy conformers within the ‘Ghemical’ force field.[229]

Machine learning techniques

Decision tree induction

See Section 4.3.1.

Random forest

See Section 4.3.1.4

Artificial neural networks

See Section 4.3.2

K-nearest Neighbor

See Section 4.3.5.

LAZAR

See Section 4.3.6.

Support vector machines

See Section 4.3.3.

Naive Bayes and Bayesian nets

See Section 4.3.4.

Bayesian nets represent probability distributions as directed acyclic graphs, where each node represents an attributes probability. Predictions are made by summing up probabilities for each instance. For learning the networks presented here, we used the K2 algorithm.[231]

Cross-validation

See Section 4.6.2.

All models were built with k=10, except for LAZAR where k=n (leave-one-out (LOO) cross-validation). Other means of validation are possible (e.g., holdout validation) and have been used in the past (Table 19).

Chemical similarity

See Section 3.3.1.2.

Feature reduction

See Section 4.4.

In this study, we compare several approaches:

1. Best first feature selection (BFS) using a greedy hill-climbing algorithm.[223]
2. Linear correlation analysis (CFS) by performing linear regressions for every descriptor. The nine best correlating (by measure of R^2) were selected.
3. DTI splitting criteria (DTIS) were used as the final subset. Features were taken from DTI models produced beforehand.

Quality measures

See Section 4.5

Receiver operating characteristics (ROC)

See Section 4.5

Comparison of numerical predictors and classifiers

The results of numerical models cannot be directly compared with those of the classification paradigms. One approach is the comparison of numerically predicted absorption with the classes from the original dataset by means of receiver operating characteristics (ROC) and their areas under the curve (AUC).

Results

Dataset

Drugs used for modeling and simulation cover a broad chemical range (Tanimoto coefficient: 0.702). This seems reasonable considering the dataset consists of commercially available drugs and thereby exhibit certain similarities, e.g., drug like properties. The mean (±SEM) weight of molecules within the database was 346.1 (±8.3).

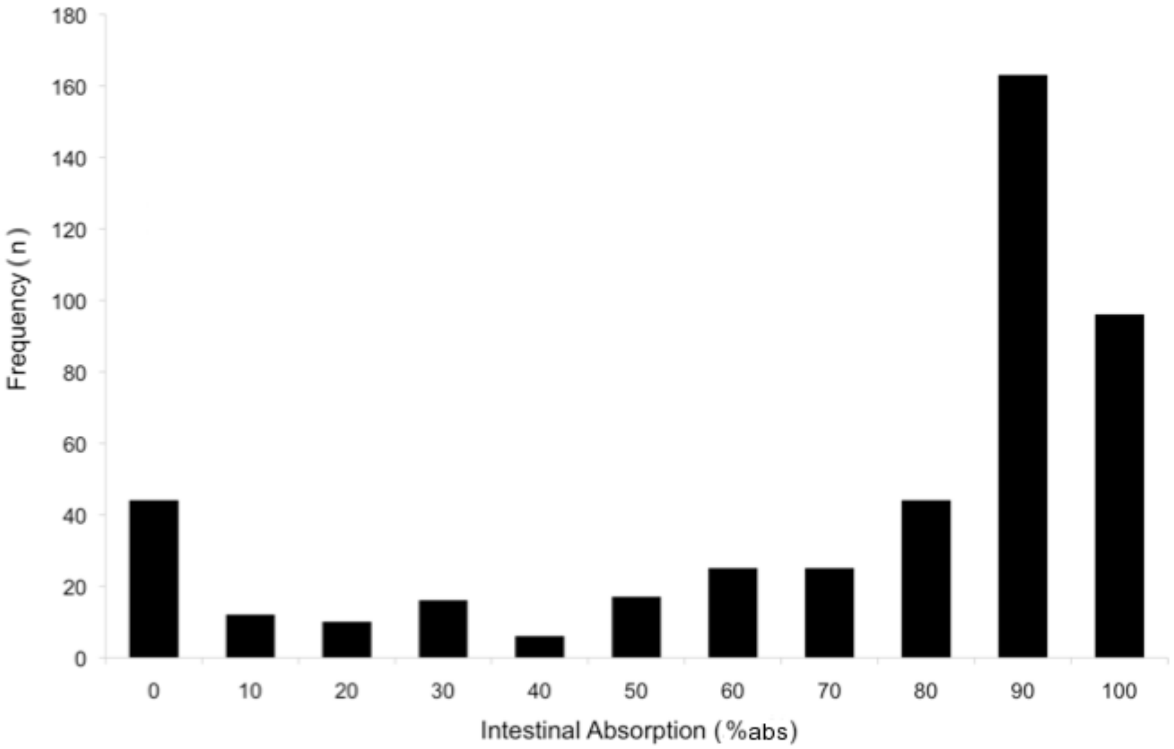


Figure 16 - Ordinal classes and corresponding absorption values (%Abs).

The present dataset exhibits a bimodal distribution with accumulation of compounds at 100%Abs and 0%Abs (Figure 16). This clearly reflects the two major routes of applications of common drugs (oral (high %Abs) or i.v. and topical administration (low %Abs)).

Feature Reduction

The descriptors selected by the different means of feature reduction are summarized in Table 14.

Best first Set	Correlation Set	DTI Set
aLogP (LogKow aLogP2)	Molar refractivity (AMR)	aLogP2 (LogKow aLogP2)
bCUTS (highest atom weighted)	bCUTS (lowest atom weighted)	H- bond donor count
H- bond donor count	H- bond acceptor count	H- bond acceptor count
H- bond acceptor count	gravitational index 4	molecular weight
gravitational index 4	gravitational index H1	longest aliphatic chain descriptor
moment of inertia descriptor (Z-axis)	gravitational index H2	tPSA
moment of inertia descriptor (XY-axis)	length over breadth descriptor (LOBMAX)	rNCS
pPSA3	pPSA2	
dPSA1	pPSA3	
dPSA3		
fNSA2		
fNSA3		
tPSA		
rHSA		

Table 14 - In order to avoid overfitting, we applied three different selection methods to identify the most relevant physicochemical features in a complete descriptor set. The best first algorithm uses a greedy hill-climbing algorithm and revealed 13 features (BFS). A linear correlation of each feature with the endpoint was performed and the nine descriptors with highest R^2 were used for modeling (CFS). For the final set, we used the seven splitting criteria revealed by Classification and Regression Trees (CART), which was produced beforehand (DTIS). For a more detailed listing of all descriptors see supporting information.

All methods selected descriptors from the charged partial surface area (cPSA) subset, partition coefficients and hydrogen bonding capacity, reflecting well-known properties of drug absorption. Strikingly, measures of molecular symmetry and mass distribution (gravitational indices, moments of inertia, longest aliphatic chains) are singled out as well. To our knowledge, compound shape is not widely used in modeling these end points.

Model Performance

Classification models

The most effective classification model in our study was built by DTI with the CHAID algorithm (CCR: 0.88, MCC: 0.75, Figure 17), followed closely by the CART algorithm (CCR: 0.84, MCC: 0.70, Figure 18).

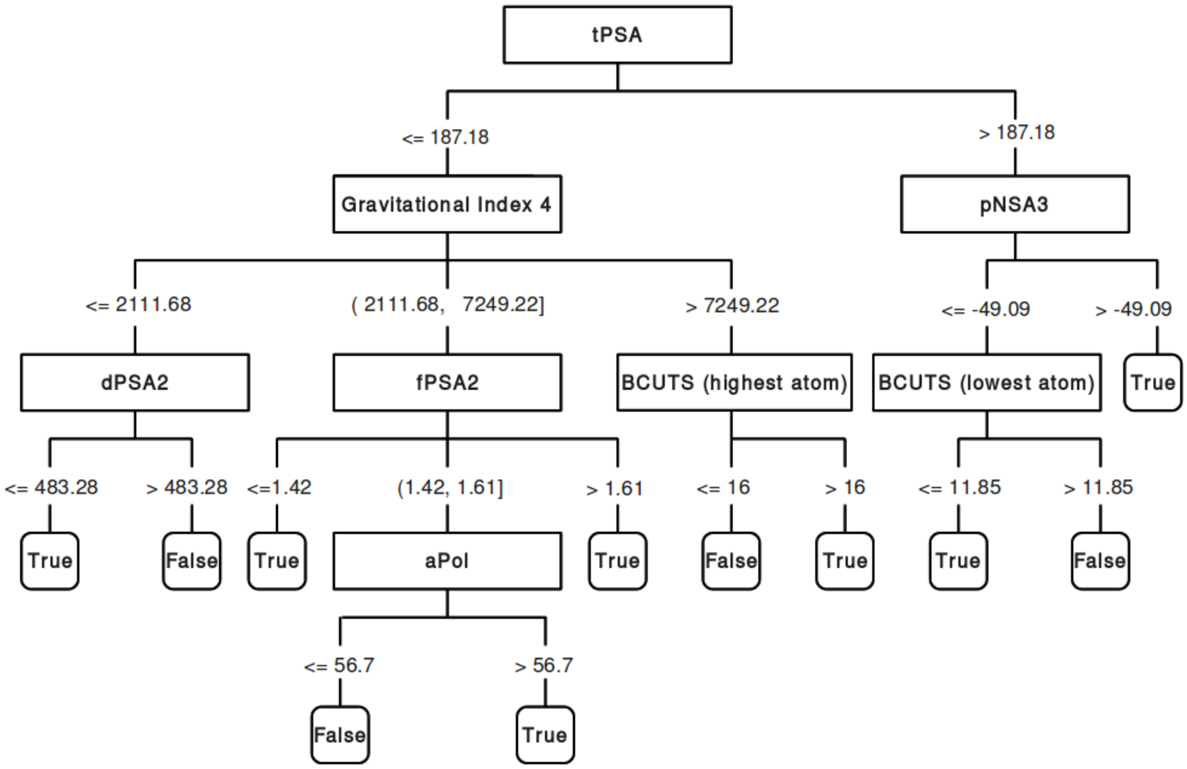


Figure 17 - Decision tree with Chi-squared interaction detector (CHAID). A maximum depth of five nodes and a minimum five cases in the parent and two cases in the child node were allowed for tree growth. Splitting criteria (boxes) and corresponding cut off values are given. Leaves are depicted as rounded boxes. Predictions achieved a corrected classification rate (CCR) of 0.88 (MCC: 0.75). The whole descriptor set was used (n = 80) for decision tree induction.

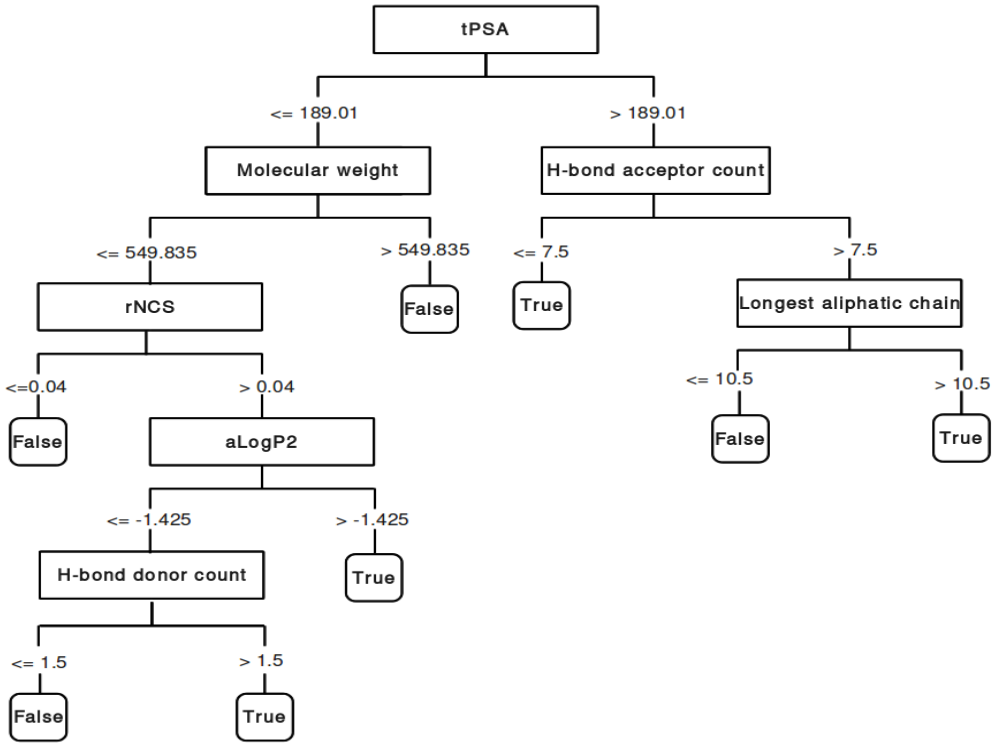


Figure 18 - Classification of human oral absorption with CART (classification and regression tree) decision tree. The 10-fold cross-validated tree performed with corrected classification rate (CCR) of 0.84 (MCC: 0.7). The Gini coefficient was used as homogeneity measure.

Of all methods applied on reduced feature sets, Bayesian techniques performed best (using the BFS subset, CCR: 0.81, MCC: 0.62). Other paradigms such as SVM did not achieve similar performance with any of the feature sets. Classification models are summarized inTable 15.

Whole Feature Set				
Method	Specificity	Sensitivity	CCR	MCC
LAZAR	0.438	0.974	0.706	0.529
CART	0.740	0.947	0.843	0.698
CHAID	0.822	0.944	**0.883**	**0.751**
Random Forest	0.589	0.947	0.768	0.583
Best first Set				
Method	Specificity	Sensitivity	CCR	MCC
SVM rbf	0.644	0.934	0.789	0.597
SVM polynomial	0.521	0.974	0.747	0.596
Multilayer Perceptron	0.534	0.974	0.754	0.607
KNN	0.534	0.974	0.754	0.607
Naive Bayes	0.685	0.934	0.809	0.629
Bayesian Nets	0.699	0.934	**0.816**	**0.639**
Correlation Set				
Method	Specificity	Sensitivity	CCR	MCC
SVM rbf	0.575	0.974	0.774	0.639
SVM polynomial	0.521	0.967	0.744	0.578
Multilayer Perceptron	0.589	0.970	0.780	**0.641**
KNN	0.603	0.931	0.767	0.558
Naive Bayes	0.521	0.931	0.726	0.492
Bayesian Nets	0.616	0.957	**0.787**	0.628
CART Set				
Method	Specificity	Sensitivity	CCR	MCC
SVM rbf	0.479	0.970	0.725	0.553
SVM polynomial	0.521	0.974	0.747	0.596
Multilayer Perceptron	0.589	0.964	0.776	0.623
KNN	0.589	0.947	0.768	0.583
Naive Bayes	0.671	0.941	0.806	0.632
Bayesian Nets	0.685	0.941	**0.813**	**0.643**

Table 15 - We used the ordinal classes "true" (≥80 %Abs) and "false" (≤ 30 %Abs) for classification. Compounds of the class “unknown" (30% < %Abs < 80%) were omitted from learning to achieve better separability for classifiers. Apart from sensitivity and specificity, highest values of corrected classification rates (CCR) and Matthews correlation coefficients (MCC) are in boldface. Results are shown for the whole dataset, best first feature selection (BFS), linear correlation analysis (CFS) and decision tree splitting criteria (DTIS).

Numerical models

The multilayer perceptron yielded the strongest numerical predictions (using the CFS subset, RMSE: 25.823, R^2: 0.600). Performance is illustrated in Figure 19.

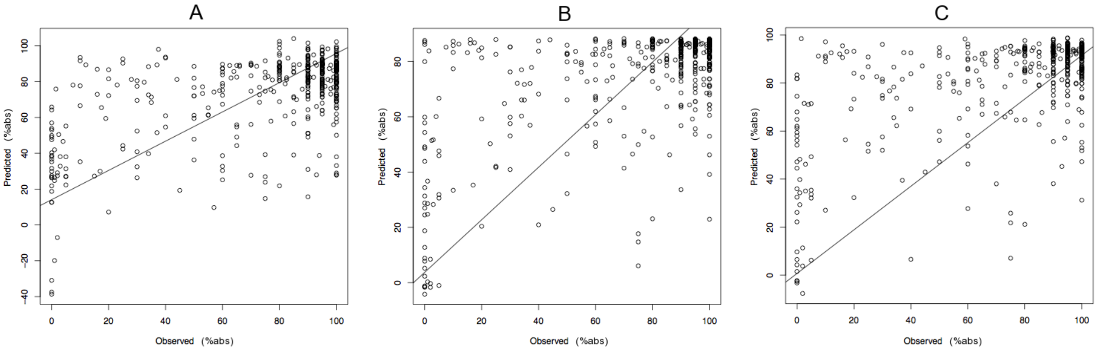


Figure 19 - Scatterplot of predicted (y-axis) vs. observed (x-axis) %abs values. Best models on numeric %abs values are given. A) Multilayer Perceptron on best first feature set (BFS) B) Multilayer Perceptron on linear correlation analysis set (CFS). C) Support vector machines with rbf kernel on decision tree splitting criteria (DTIS).

SVMs with the rbf kernel achieved comparable efficacy on the DTIS subset (RMSE of 26.953; R^2: 0.590). Other methods did not perform as well (Table 16).

Best first Set		
Method	CC	RMSE
SVM rbf	0.574	27.648
SVM polynomial	0.561	27.807
Multilayer Perceptron	**0.590**	**26.390**
Correlation Set		
Method	CC	RMSE
SVM rbf	0.559	27.828
SVM polynomial	0.546	28.535
Multilayer Perceptron	**0.600**	**25.823**
DTI Set		
Method	CC	RMSE
SVM rbf	**0.590**	26.953
SVM polynomial	0.544	28.773
Multilayer Perceptron	0.588	**26.099**

Table 16 - Prediction of %Abs was assessed using Support vector machines with kernels and multilayer perceptron. Methods were applied on reduced feature sets: best first feature selection (BFS), linear correlation analysis (CFS) and decision tree splitting criteria (DTIS). As quality measures root mean squared error (RMSE) and correlation coefficient (R2) are given. Best results are given in boldface.

Recoding of numerical predictions into classes

As a means of salvaging predictive power from the rather mediocre numerical models, we attempted to recode the predicted %Abs values into classes based on a retrospective analysis using ROC curves (see Table 17 and Figure 20)

Threshold 30	**Best first Set**				
Method	Specificity	Sensitivity	CCR	MCC	AUC
SVM rbf	0.164	0.875	**0.519**	0.590	0.786
SVM polynomial	0.137	0.835	0.486	0.546	0.786
Multilayer Perceptron	0.329	0.686	0.508	**0.715**	0.746
	Correlation Set				
Method	Specificity	Sensitivity	CCR	MCC	AUC
SVM rbf	0.205	0.818	**0.512**	0.586	0.755
SVM polynomial	0.110	0.871	0.490	0.496	0.767
Multilayer Perceptron	0.301	0.719	0.510	**0.655**	0.757
	CART Set				
Method	Specificity	Sensitivity	CCR	MCC	AUC
SVM rbf	0.205	0.871	**0.538**	0.579	0.780
SVM polynomial	0.110	0.875	0.492	0.479	0.773
Multilayer Perceptron	0.205	0.637	0.421	**0.681**	0.774

Table 17 - Performance of numeric models is shown after recoding into classification scale. We applied cut-off values from initially set ordinal classes. For performance measurement only positive class (≥ 80 %Abs) and negative class (≤ 30 %Abs) was used. Compounds classified as unknown were omitted. The corrected classification rate (CCR), Matthews correlation coefficient (MCC), specificity, sensitivity and area under the ROC curve (AUC) are indicated for each model. Best results for coefficients are given in boldface for all reduced feature sets: best first feature selection (BFS), linear correlation analysis (CFS) and decision tree splitting criteria (DTIS).

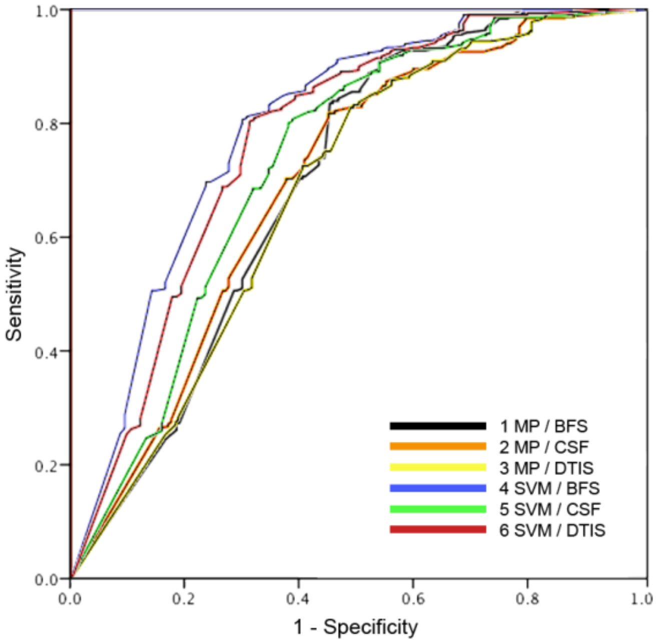


Figure 20 - ROC curves are shown of best recoded models according to highest achieved corrected classification rates (CCR) and Matthews correlation coefficients (MCC). Multilayer perceptron on best first feature set (BFS) (1), linear correlation analysis set (CFS) (2) and decision tree-splitting criteria (DTIS) (3). Support vector machines with rbf kernel on BFS (4), CFS (5) and DTIS (6).

For each model, optimal thresholds were selected by determining the one with the highest Youden index (J). Instances were recoded into the two-class case according to these thresholds. Models and their performance after recoding are summarized in Table 18. The SVM model with the rbf kernel was most precise (CCR: 0.72; MCC: 0.47, using the BFS subset), outperforming the MP model.

Best first Set					
Method	Th_{opt}	Specificity	Sensitivity	CCR	MCC
SVM rbf	82.0	0.541	0.836	0.689	0.398
SVM polynomial	79.6	0.613	0.842	0.727	**0.464**
Multilayer Perceptron	78.3	0.723	0.739	**0.731**	0.443
Correlation Set					
Method	Th_{opt}	Specificity	Sensitivity	CCR	MCC
SVM rbf	78.9	0.574	0.842	0.708	0.430
SVM polynomial	78.0	0.526	0.901	0.714	**0.471**
Multilayer Perceptron	68.4	0.538	0.896	**0.717**	0.468
CART Set					
Method	Th_{opt}	Specificity	Sensitivity	CCR	MCC
SVM rbf	83.9	0.570	0.809	0.689	0.391
SVM polynomial	83.4	0.602	0.754	0.678	0.358
Multilayer Perceptron	72.3	0.619	0.840	**0.729**	**0.461**

Table 18 - Optimal cut points for thresholds were determined using the maximization of the Youden indices in receiver operating characteristics analysis. Specificity, sensitivity, CCR and MCC for all models are indicated at the optimal threshold (Th_{opt}). The best results are given in boldface for all reduced feature sets: best first feature selection (BFS), linear correlation analysis (CFS) and decision tree splitting criteria (DTIS).

Discussion

Dataset

Our dataset of 458 substances covers a broad range of small molecule drugs as indicated by the high value of dissimilarity within the descriptor space employed. The distribution of absorption ratios is bimodal with a small peak at the low end of the spectrum and a larger one for highly absorbed molecules (Figure 16). This reflects the desire to bring to market orally administrable drugs. Models based on this dataset should therefore best be applied in late-stage drug development. Even though the dataset is unbalanced, the sensitivity of our models is not impaired.

Wang et al. [232] proposed to use drug subsets with similar pharmacological targets when modeling human intestinal absorption. While this approach may be of use in late-stage optimization, we feel that general physiological features cannot be deduced when examining such restricted data sets.

Our models intentionally disregard mechanistic minutiae of intestinal absorption (e.g., transcellular versus paracellular pathways) as we are predicting the final endpoint of human intestinal absorption and not specific pathways. Members of our group have shown the validity of this approach for even more complex endpoints.[233] Furthermore, the original data source does not provide sufficient information on the specific absorption kinetics and metabolism on the level of the intestinal epithelium.

Feature Selection

All feature sets include descriptors of PSA. Palm et al. demonstrated its correlation with human intestinal absorption and the CACO-2 cell model.[234] In models of Winiwarter et al. [235], polar surface area (PSA) was emphasized as one of the most important parameters to predict drug permeability. However, it was reported that an excellent sigmoidal relationship with high correlation could be established between the absorbed fraction after oral drug administration to humans and PSA.[236] In line with other groups [237, 238] we clearly disapprove of this approach. As a single feature, PSA is not a reliable criterion to distinguish poorly absorbed from well-absorbed compounds. Seven descriptors of the cPSA set appear in the BFS set of features. It is important to note that while all of these concern charge analysis, they are distinctly different. Nonetheless, we performed an additional analysis and found low intercorrelation between these features (r_{avg} 0.59 ± 0.06 SEM). Highest rsig (0.90) is seen for dPSA3 and fNSA2, both of which are derived from central descriptors of the cPSA set. Specifically, fNSA2 puts charge into relation with molecular topology while dPSA3 weighs positive against negative charge contributions. Therefore, both contribute distinct molecular information to the models and hence have not been removed from the final dataset. Additionally, any intercorrelation is penalized by cross-validation and does not introduce an overfitting bias.

According to current understanding, the feature selection paradigms singled out descriptors of lipophilicity, charge (e.g., aPol), hydrogen bonding descriptors, and molecular weight (as selected in CART DTI trees). These features are already known from studies of human jejunal permeability ($LogP_{eff}$) [18, 235] and deconvolution studies of human absorption rate constants.[239] Zhao et al. further demonstrated that hydrogen bonding is the rate-limiting step in absorption kinetics.[240]

Repeated inclusion of gravitational Indices (CHAID, BFS, CFS), moments of inertia (BFS), length over breadth (CFS), and longest aliphatic chain (BFS, CART) indicate the importance of molecular mass distribution and geometry in modeling oral absorption. This seems reasonable as smaller molecules have better passive permeation capability than compounds with long aliphatic motifs. Moreover, BCUT descriptors [241] were selected by two paradigms (CHAID, BFS). These features are defined as eigenvalues of modified connectivity matrices with frequent application in drug discovery [242]. Their discriminatory power for aqueous solubility is well known [243] and therefore confirms the importance of this physicochemical property in absorption kinetics.

Best first feature reduction and linear correlation are two commonly used means of reducing dimensionality of the descriptor space. The use of features selected from DTI models learned from the same data is rather unusual. We consider this a valid approach in that the feature set provides a mechanistical theory, which the models created a posteriori seek to verify. There is no unreasonable flow or leakage of information into the learning process (as in an overfitting bias) compared to reducing features using the other paradigms.

Individual Models

The DTI algorithms provided the strongest models (Table 15). Other paradigms, such as SVM classifiers, showed far weaker performance. These observations are in line with other studies.[177, 233] DTI often outperforms other machine learning methods when moderately sized and skewed datasets are used. Table 19 gives a summary of recent modeling attempts. Comparable performance is achieved only by work based on DTI and Gaussian kernels such as Obrezanova et al. [244] who, however, fail to cross-validate their models. In terms of accuracy, our models (DTI) are only rivaled by work by Shen [245], which, again is not cross-validated. Remarkably, many studies choose not to employ cross-validation, resulting in accuracy measures which overestimate their power in unseen data. Hou et al. [246] reported very strong models in a comparable context using SVM. Predictions achieved MCC of up to 0.89. Their dataset, however, had been artificially expanded by extrapolating %Abs values from bioavailability data. Moreover, the inclusion of redundant descriptors may have lead to overfitting, as has been reported previously [238].

It is worth noting that the PSA descriptors of CDK also contain the fragment-based method (tPSA) introduced by Ertl et al..[247] Inclusion of different PSA paradigms does not introduce redundancy. The method established by Ertl et al. considers molecule fragments, which might only be exposed to the environment when drugs are dissolved in the aqueous intestinal lumen. Other implementations focus on charge and total molecular surface area.

On reduced features, Naive Bayes and Bayesian nets performed well. Their impressive results seem at first glance surprising, especially as these paradigms assume independence of variable. Although this is rarely given in real world data, dependence can be minimized by eliminating redundant and therefore non-independent features. It can be argued that with the splitting of data into test and training sets, the independence bias is not equally distributed over both sets. Predictions of unseen data should then be interpreted with caution.[218] This assumption holds less well in the case of randomly performed cross-validation. The quality measures presented here might therefore give a more realistic estimate of the predictive power. Provided that the compound of interest fits the chemical space analyzed in this study, Naive Bayes models should classify it correctly.

Work	N	CV	Paradigm	R2	RSME	CCR	MCC	Acc	Sens.	Spec.
Zhao et al., 2001[248]	241	No	Regression	0.74	14	-	-	-	0.95	0.72
Niwa et al., 2003[249]	86	No	ANN (general regression)	-	22.8	-	-	-	-	-
Niwa et al., 2003[249]	86	No	ANN (probabilistic)	-	-	0.75	0.612	80%	1	0.5
Bai et al., 2004[250]	1260	No	DTI (CART)	-	-	-	-	79 - 86%	-	-
Liu et al., 2005[251]	169	Yes	SVM (gaussian kernel)	0.73	14.08	-	-	-	0.98	0.66
Jones et al., 2005[252]	241	No	Kernel	-	22%	-	-	-	0.9	0.46
Deconinck et al.,2005[253]	141	Yes	DTI (CART)	-	-	-	-	65%	0.89	0
Iyer et al., 2006[254]	188	No	Membrane-Interaction QSAR	0.68	-	-	-	-	-	-
Hou et al., 2007[238]	455	No	Genetic programming	-	-	-	0.836	-	1	0.64
Yan et al., 2008[255]	552	No	PLS	0.83	18.18	-	-	-	-	-
Yan et al., 2008[255]	552	No	SVM (rbf kernel)	0.89	16.53	-	-	-	-	-
Reynolds et al., 2009[256]	567	No	Nonlinear regression	0.84	35	-	-	-	-	-
Obrezanova et al., 2010[244]	260	No	Kernel	-	-	-	-	91%	-	-
Obrezanova et al., 2010[244]	260	No	DTI (unspecified algorithm)	-	-	-	-	85%	-	-
Shen et al., 2010[245]	578	No	SVM (polynomial)	-	-	0.928	0.909	98%	0.998	0.859
Shen et al., 2010[245]	578	No	SVM (rbf)	-	-	0.948	0.932	98%	0.998	0.897
Guerra et al., 2010[257]	202	Yes	ANN	-	-	-	-	73%	-	-
Suenderhauf et al., 2010	458	Yes	DTI (CHAID)	-	-	0.883	0.752	92%	0.944	0.882
Suenderhauf et al., 2010	458	Yes	DTI (CART)	-	-	0.843	0.698	91%	0.947	0.740
Suenderhauf et al., 2010	458	Yes	ANN (numeric)	0.6	25.823	-	-	-	-	-
Suenderhauf et al., 2010	458	Yes	ANN (recoded)	-	-	0.717	0.468	79%	0.896	0.538

Table 19 - Representative models for human intestinal absorption are summarized. Indicating size of dataset used (n), the use of cross-validation (CV), paradigm and algorithm used and performance estimates (coefficient of determination [R2], root mean squared error [RMSE], corrected classification

rate [CCR], Matthews correlation coefficient [MCC], accuracy [Acc], sensitivity and specificity). Accuracy was calculated as true hits (true positives and true negatives) divided by n.

The numeric models presented here exhibit low predictive power as visualized in Figure 19. This might be caused by the bimodal distribution of the dataset. The clustering of compounds around low and high levels of absorption reflects the two major galenic classes of drug: orally and intravenously administered compounds. Hence, the instance space is not entirely covered (Figure 16). Regression models are likely to perform badly on such data. Indeed, the achieved R^2 values range from 0.544 to 0.600, confirming that linear regression models are a completely inappropriate method type for the present dataset. This holds even in the case where compounds are grouped together (such that a group has similar activities), because of the linear model's constant slope. It would be more appropriate to perform binning of instances into two classes and analyze them by classification. Because focusing on two classes can improve numeric models, we stress the importance of choosing appropriate algorithms for the dataset at hand.

It also should be noted that we succeeded in producing models of high accuracy (up to 92% with DTI) without specifically incorporating the influence of efflux transporters such as P-gp. This indicates that these are not major influencing factors. The poorer performance of numerical models therefore seems to be intrinsic to the paradigms.

Numeric vs. Classification

Comparison of classification and numeric models is not straightforward. Numeric measures of accuracy are RMSE or R^2. In a classification system, confusion matrices and corresponding quality measures (CCR, MCC) are used. Both model types may therefore not be compared directly. In an attempt to do so indirectly, we translated numeric predictions into a classification scale. As expected, performance of numeric models was worse compared to genuine classification. We analyzed predictions with ROC graphs (Figure 20) to estimate predictive power for the well-absorbed class. Models achieved a reasonable sensitivity, which is reflected in high AUC values (Table 17). In other words, numerical models tended to generally overestimate absorption ratios. By determining optimal cut-off values using the best Youden Index, we markedly improved models in terms of specificity. Treated in this fashion, we state that numeric and classification models perform comparatively strong (Table 15, Table 18)

Conclusion

While intestinal absorption of drugs in humans is mostly governed by passive diffusion, it is potentially influenced by several other factors. Our models can be seen as a combined endpoint analysis as they disregard, among other aspects, the location of absorption and specifics of administration (e.g., galenics, counterions). Performance, however, is not impeded by these generalizations. The dataset comprises the entire range of absorption ratios and has a great chemical diversity in the descriptor space employed.

Although we used differing approaches to reduce features, certain descriptors were present in all sets. Descriptors of charge and lipophilicity reflect the current understanding of drug absorption in humans. Our models show the importance of molecular shape and complexity on absorption. Small size, little branching, and equal distribution of mass (as represented by descriptors of the moments of inertia) seem to be of advantage in oral absorption.

The advantages of computational methods in the prediction of oral absorption have been described previously, e.g., Norinder et al.[15] In clinical practice, drugs fall into just two categories with little overlap: those which are orally administrable and those for which other routes of administration have to be taken (for example intravenous injection or topical application). Therefore, specific numerical values, such as absorption ratios are often considered to be of less importance than classification.

Acknowledgments

Andreas Maunz and Christoph Helma ware supported by the EU FP7 Project OpenTox (Contract Number Health-F5-2008-200787).
Claudia Suenderhauf was supported by the Swiss National Foundation (Grant No. 323530-119218).

5.2 New Computational Models for Predicting Drug Brain Penetration

Claudia Suenderhauf, Felix Hammann, and Jörg Huwyler

Pharmaceutical Technology
Department of Pharmaceutical Sciences, University of Basel, Klingelbergstrasse 50,
CH-4056 Basel, Switzerland

Submitted to Journal of Medicinal Chemistry

Abstract

Drug penetration into the central nervous system is dependent on a molecule's ability to cross the blood brain barrier (BBB). Limited passive diffusion and active efflux and influx systems account for the complexity of this highly regulated process. It was our aim establishing models of drug permeation including both active and passive transport systems.

We collected a database of 163 compounds where information on the *in vivo* surface permeability product (LogPS) in rats was available. We used the DRAGON toolkit and the Chemical Development Kit (CDK) to calculate physicochemical descriptors. Decision trees were induced on both descriptor sets. We were able to establish models with corrected classification rates (CCR) of 90.9% - 93.9%. An Ant colony optimization (ACO) based binary classifier was used to search for the most predictive chemical substructures. The best model yielded a CCR of 89%.

Decision trees revealed descriptors of lipophilicity (partition coefficient) and charge (polar surface area), which were also described in models of passive diffusion. However, measures of molecular geometry and connectivity could be related to an active drug transport component.

Introduction

See Section 3.2.1.2.

Development of new central nervous system (CNS) active drugs is hampered by limited brain permeation. As invasive methods have proven themselves to be ineffective and risky for patients, the systemic application is the preferred route for drug administration into the brain.[28, 29] Hence, blood brain barrier (BBB) permeability is a feature absolutely mandatory for any drug which targets the CNS. It is desirable to have estimates on a compounds behavior on level of the BBB as early as possible in the drug development process.

The process of passive brain permeation is well characterized and accurate computational models to predict molecule behavior exist.[33-36] Major physicochemical determinants are lipophilicity, molecular weight, and measures of molecular polarity.[32] However, such expert-based rules do not accurately reflect the complexity of interactions as they disregard the biochemical processes mediated by transport proteins.[37]

Computational models to discriminate substrates, inhibitors, and inducers of P-glycoprotein (P-gp) have successfully been established.[258] It is tempting from a mechanistic point of view to join the various models for passive diffusion and active transport to predict brain penetration in general. However, every model suffers from a varying degree of uncertainty which accumulates the more of them are concatenated. To avoid such shortcomings, it is more efficient to generate a single model of a complex phenomenon rather than sequentially apply model after model. Members of our group have shown the feasibility and validity of this approach.[259, 260]

We applied modern machine learning algorithms to predict this highly complex endpoint. We assembled 163 *in vivo* BBB permeability-surface area (PS) product (LogPS) experiments from literature. LogPS values are usually calculated from internal carotid artery perfusion studies in rats. This procedure is considered superior to other methods like blood/brain partition measurement at steady state (LogBB), as it lacks systemic distribution effects, which distort brain penetration substantially.[51] The majority of high quality data found in literature was gathered in rats. We decided to omit data acquired in other species to avoid interspecies variability and reduce noise in our models.

Materials & Methods

Dataset

We assembled a dataset of 163 small molecules from literature where information on *in vivo* BBB permeability-surface area (PS) product, usually given as its logarithm (LogPS), was available.[51, 261-273] Besides wild type animal data, we found also LogPS values from transgenic rats. Single transporter knockout animals are particularly useful when active transport mechanisms are studied. However for the present work we exclusively used data from wild type animals, as we aimed to model brain penetration entirely under physiological conditions (i.e. including also active transport).

The paradigms used in the present study were classification algorithms. We therefore aimed to split data in two classes, according to cut-off values published literature.[274, 275] LogPS values >= -2 were judged as readily penetrating the brain and received the label "CNSp+" (n = 70), while measures <= -3 were labeled "CNSp-" (n = 61). To increase discriminatory power of our models, values between -2.1 and -2.9 were exempt from classification learning (n = 32). The final dataset used consisted therefore of 131 compounds.

Structural information was retrieved from the National Library of Health database PubChem (http://pubchem.ncbi.nlm.nih.gov/). For salts, we removed the counterion prior to further considerations. Conversion to three-dimensional structure representation was achieved by using lowest energy conformers within the Ghemical force field.[229]

Physicochemical descriptors

See Section 4.2 and 4.6.3.

Fingerprints

See Section 4.1.3.

Machine learning techniques

Decision tree induction

See Section 4.3.1.
CHAID and CART were grown to a maximum depth of 3 and 5 nodes, respectively. We set minimum cases for parent nodes to 10 instances and allowed 5 cases in the child nodes.

Ant colony optimization

Ant colony optimization is a natural computing paradigm introduced by Dorigo et al.[224] The algorithm uses an abstraction of ants foraging behavior to select meaningful features. Higher-dimensional QSAR studies, e.g., ligand docking, routinely apply ACO alongside other optimization paradigms. With a few modifications, ACO can be used as a feature selector, i.e. to identify attributes that carry information on the endpoint of interest. For this study, we applied a variant of ACO algorithm recently published by our group (Section 5.5).

Cross-validation

See Section 4.6.

Chemical Similarity

See Section 3.3.2.

Quality measures

See Section 4.5.

Software used

See Section 4.6.3.

Results

Chemical Similarity

The Tanimoto coefficient for our dataset (n = 131) was 0.282. This great dissimilarity indicates that the data span a reasonable chemical space.

DTI performance

Decision trees using CHAID algorithm trained with Dragon descriptors yielded the best results (Figure 21). This tree classified compounds with a CCR of 93.9% (MCC: 87.9%).

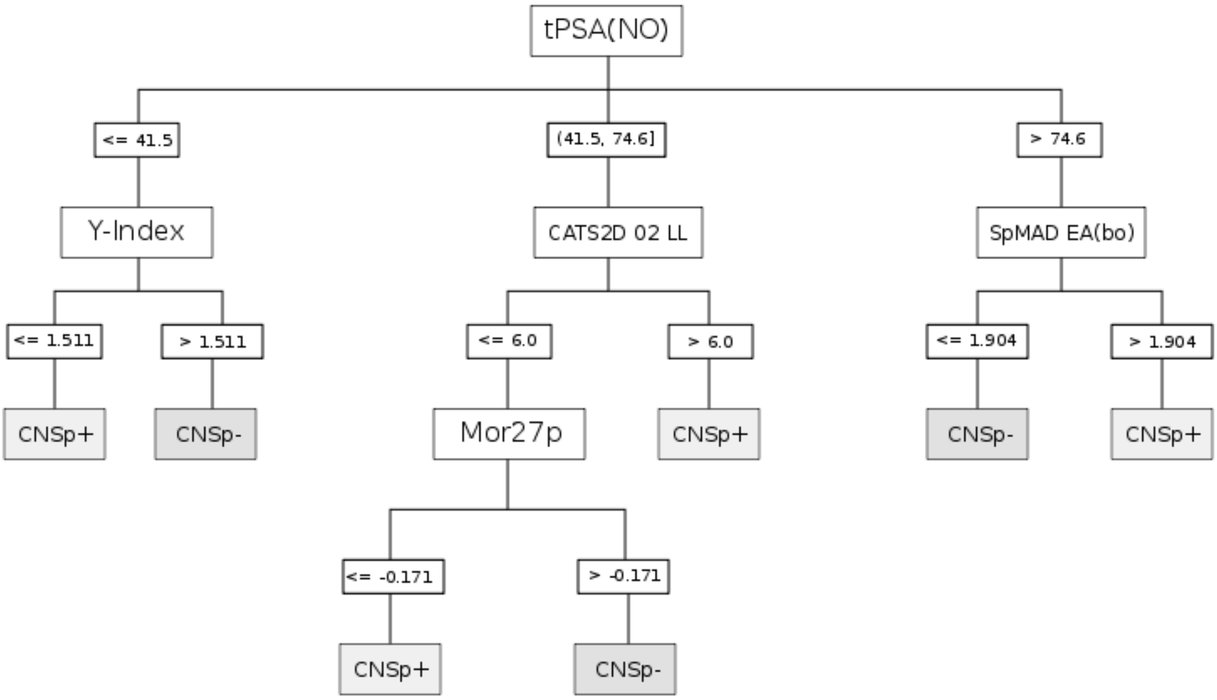


Figure 21 - Chi-squared automatic interaction detector (CHAID) was trained on the whole feature set provided by DRAGON toolkit (n= 4885). The resulting model performed with a corrected classification rate of 93.9%. "CNSp+" indicates good permeation and "CNSp-" stands for bad brain permeation.

Features were topological polar surface area (tPSA[NO]) derived from polar fragments (i.e., oxygen and nitrogen), the Balaban Y Index (Y-Index), the distance of lipophilic pharmacophore groups (CATS2D 02 LL), 3DMoRSE descriptor weighted by polarizability (Mor27p), and spectral mean absolute deviation from the edge adjacency matrix, weighted by bond order (SpMAD EA [bo]). When trained with CDK descriptors, the paradigm performed slightly worse, achieving a CCR of 90.9% (MCC: 81.7%) (Figure 22).

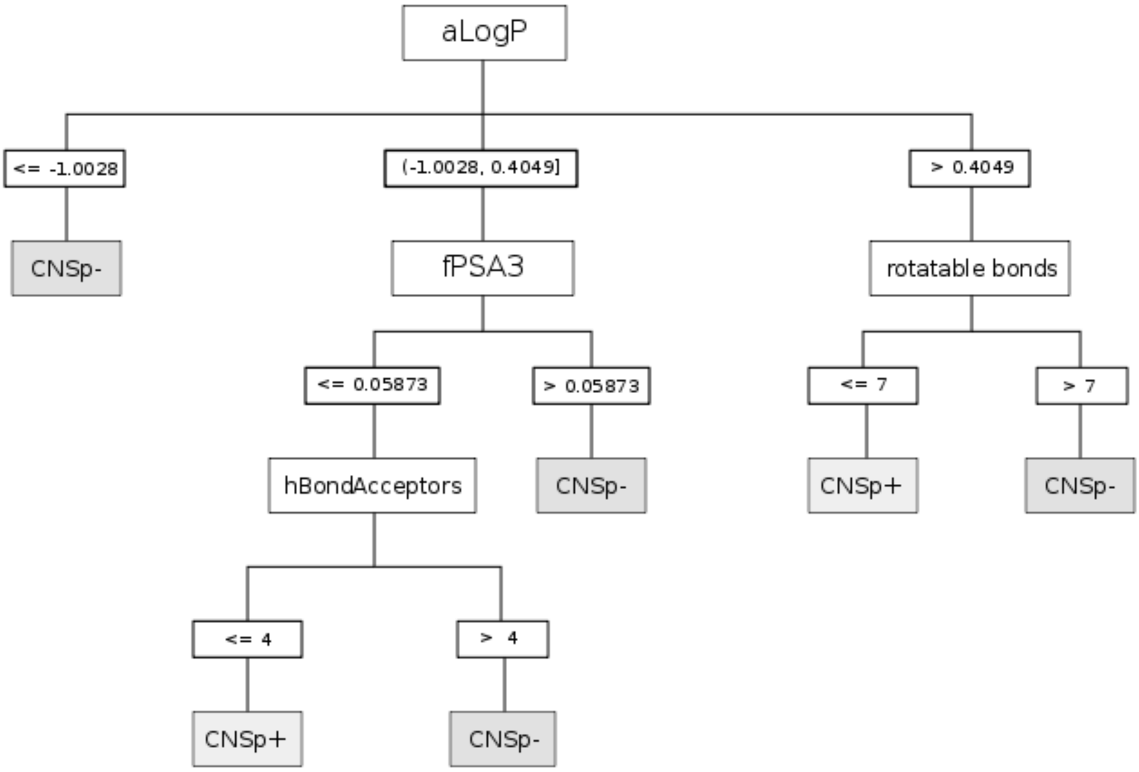


Figure 22 - The tree built by Chi-squared automatic interaction detector (CHAID) on CDK descriptors (n= 81) is shown. The cross-validated model achieved a corrected classification rate of 90.9%. When a molecule reaches a leave indicating "CNSp+" it is judged to exhibit good brain permeation. When "CNSp-" is reached, the molecule will not pass the BBB.

Splitting criteria were partition coefficient (aLogP), charge weighted partial positive surface area divided by total molecular surface area (fPSA), hydrogen bond acceptor count (hBondAcceptors), and rotatable bonds count (rotatable bonds). Models created with CART paradigm could not match the performance of CHAID. Trained on DRAGON and CDK feature sets it yielded a CCR of 90.8% (MCC: 81.6%) and CCR of 89.8% (MCC: 79.9%), respectively. Trees are summarized in Figure 24 and Figure 26.

.

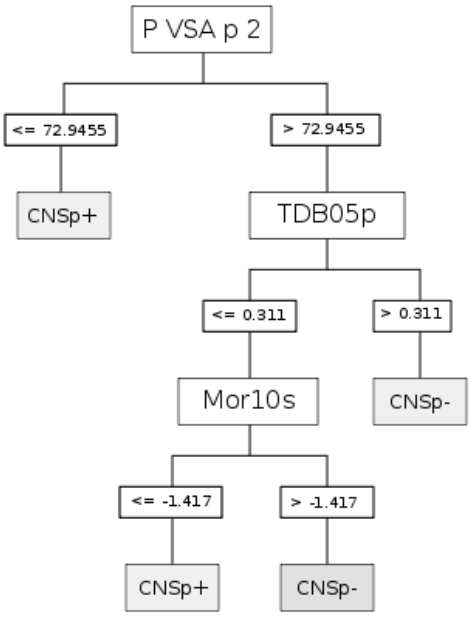


Figure 23 - Classification and regression tree (CART) on DRAGON descriptors (n= 4885). The corrected classification rate was 90.8%.

When trained with Dragon descriptors, CART used the amount of van der Waals volume having polarizability over one (P_VSA_p2), three-dimensional (3D) autocorrelation weighted for polarizability (TDB05p), and 3D-MoRSE descriptor, weighted by ionization potential (Mor10s). Partition coefficient (aLogP), topological polar surface area (tPSA), and highest eigenvalue weighted for the lowest atom in the Burden matrix (BCUTS) were chosen. Table 20 gives a comprehensive summary of model performance. Features used for classification are given in Table 21.

	DRAGON		CDK	
	CHAID	CART	CHAID	CART
CCR	93.9	90.8	90.9	89.8
MCC	87.7	81.6	81.7	79.9
Spec	93.4	90.2	91.8	86.7
Sens	94.3	91.4	90	92.9

Table 20 - Performance of chi squared automatic interaction detector (CHAID) and classification and regression tree (CART) on Chemical Development Kit (CDK) and DRAGON descriptors is summarized. Corrected classification rate (CCR), Matthews correlation coefficient (MCC), Specificity (Spec), and Sensitivity (Sens) are given. All models presented were cross-validated and quality measures indicate a realistic performance estimate for unseen data.

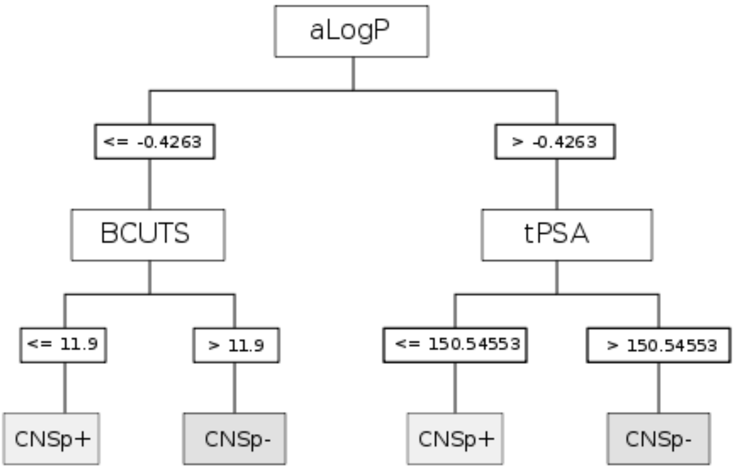


Figure 24 - We summarized the result of classification and regression tree (CART) on CDK descriptors (n= 81). The predictive power of the cross-validated model was 89.8%.

Descriptor sets	Paradigm	Splitting criteria	Comment
DRAGON	CHAID	tPSA(NO)	Topological polar surface area (only considering Nitrogen and Oxygen)
		Yindex	Balaban Y index
		CATS2D_02_LL	CATS2D descriptor lipophillic-lipophillic at lag 02
		Mor27p	3D-MoRSE descriptor, weighted by polarizability (signal 27)
		SpMAD_EA(bo)	Spectral mean absolute deviation from the edge adjacency matrix, weighted by bond order
	CART	P_VSA_p_2	P_VSA-like descriptor, weighted on polarizability (bin2)
		TDB05p	3D topological distance based descriptors- lag 5 weighted by polarizability
		Mor10s	3D-MoRSE descriptor, weighted by I-state (signal 10)
CDK	CART	aLogP	Partition coefficient as defined by Ghose-Crippen
		BCUTS	The number of highest eigenvalue, weighted for the lowest atom
		tPSA	Topological polar surface area
	CHAID	aLogP	Partition coefficient as defined by Ghose-Crippen
		fPSA3	Charge weighted partial negative surface area/ total molecular surface area
		hBondAcceptors	Hydrogen bond acceptor count
		rotatable bonds	Rotatable bonds count

Table 21 - Features revealed by DTI to predict brain permeation are shown. Descriptor sets, paradigm, and its selected criteria as well as a short explanation of the descriptor is given.

ACO performance

The best performing subset of fingerprints revealed by ACO is summarized in Table 22. This subset of chemical substructures performed with a CCR of 82% (MCC: 64%). Figure 27 shows the ROC curve and cutoff point (circle). The corresponding area under the curve (AUC) was 0.89.

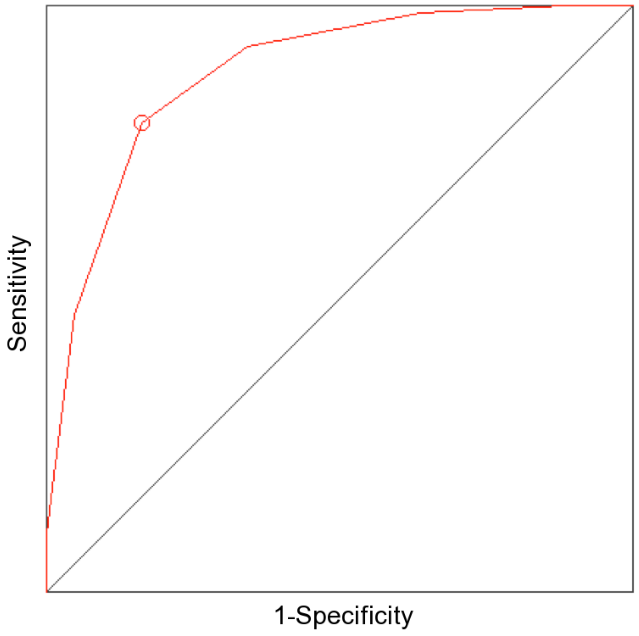


Figure 25 - The ROC curve is depicted corresponding to the best fingerprint set revealed by ant colony optimization (ACO). Cut off value is indicated by red circle. Fingerprints were selected from the MACCS key set (n= 166). This subset consisted of nine fingerprints and achieved a corrected classification rate of 82%.

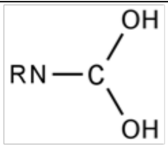
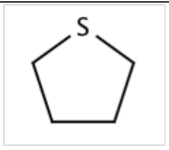
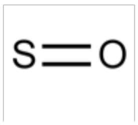
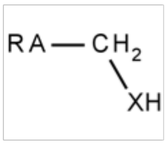
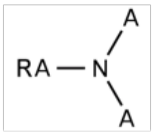
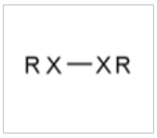
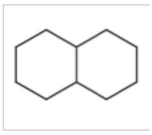
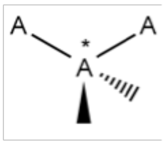
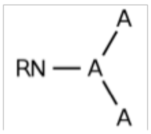
No	Sample Structure	SMARTS	Descripton
23		[#7]~[#6](~[#8])~[#8]	Nitogen connected to carbon atom, which is connected to two Oxigen atoms.
36		[#16R]	Any heterocycle containing a sulfur atom.
60		[#16]=[#8]	Oxigen and Sulfur connected by a double bond.
82		*~[CH2]~[!#6;!#1;!H0	Any atom connected to CH2, which is itself connected to a heteroatom with at least one hydrogen atom.
122		*~[#7](~*)~*	Any atom connected to Nitrogen. Nitrogen has to be connected with any two additional atoms.
130		[!#6;!#1]~[!#6;!#1]	Two heteroatoms connected to each other.
145		*1~*~*~*~*~*~1	Six ring structure, occuring twice in molecule (They do not have to be directly connented).
150		*!@*@*!@*	One intramolecular chirality centre.
156		[#7]~*(~*)~*	Nitrogen connected to any three atoms.

Table 22 - The fingerprints selected from the MACCS keys are given with their internal number (No), SMART code, and a short explanation of the substructure. In the sample structure, A stands for any atom, X for a heteroatom, and R any molecular rest.

Discussion

Dataset

The low level of chemical similarity (Tanimoto coefficient = 0.282) reflects the broad chemical space covered by our dataset. We restricted the present dataset to results of experiment in rats. For this reason, our dataset seems at first glance smaller compared to other work published in the field. We feel that artificially expanding datasets by mixing data from different species would introduce noise into deduced models. Moreover, we only used results from experiments conducted in wild type animals in order to not intentionally exclude any actively transported compounds. Active transport plays a major role in BBB permeation and can alter pharmacokinetics of a drug substantially.[276] Moreover, one can hardly assure purity of a dataset including only passively transported molecules. The characterization of active transport mechanisms is still an ongoing topic of research and active transport mediated by yet unknown transporters could remain undetected when saturation occurs at very low concentrations.

In the past, it was criticized that binning in CNS positive and CNS negative substances referred to presence or absence of pharmacological CNS activity, respectively.[277] We met these justified concerns in our considerations. Pharmacological activity is a qualitative and inadequate measure for brain permeation capacity and it is advisable to use a quantitative permeability measure like LogPS for classification instead. In the present study, the distinction in positively and negatively classified molecules refers to compounds with LogPS values <=-2mg/ml/s and >=-3mg/ml/s, respectively.

Decision tree models

One of the main advantages of DTI is the human-readable output they produce. Our models did not only predict this highly combined endpoint with excellent performance, but also gave insights into the biological processes involved. Interestingly, some features revealed were already used in models of passive brain permeation. Descriptors of lipophilicity and charge are frequently used to predict membrane permeation. It is therefore not surprising that three out of four paradigms selected partition coefficient (aLogP), the distance of lipophilic pharmacophore groups (CATS2D 02 LL) and/or polar surface area (PSA) as splitting criteria.

When we compared trees using CDK descriptors, we found that both paradigms set a much lower threshold for splitting on aLogP than earlier defined rules do.[33, 278-280] This could be an indication for active transport involvement. Recent studies refer to increasing lipophilicity as a major rate-limiting feature for P-gp interactions and played a predominant role in DTI models predicting P-gp inhibitors and substrates.[258, 281] We could therefore not confirm the opinion that high lipophilicity would be generally associated with good brain permeation.[179] While we found that it was clearly an important feature to split data on, aLogP unfolded its predictive power for the present endpoint only in combination with other descriptors.

Polar surface area was present in virtually all models. Other groups observed a corresponding role of this feature in CNS penetration.[179] Generally, our models revealed that polar molecules (PSA) were associated with bad BBB permeation. The cutoff value for classifications varied substantially between our models, but generally spoken higher molecular polarity hindered passage into the hydrophobic milieu of the brain endothelial cells. The tree grown by CHAID with Dragon descriptors used tPSA as the root splitting criterion. Although earlier work implies that PSA values over 60-90 Å^2 are generally associated with bad brain permeation, the paradigm detected in combination with connectivity measures positive instances.[278, 282, 283] This is in line with other work, where high capacity for polar interaction and low connectivity were crucial features of P-gp substrates.[258, 284]

We observed that CHAID attributed good BBB permeation for compounds with less than four hydrogen bond acceptors. This is an interesting finding as it is generally agreed that hydrogen bond acceptors are less confining for passive diffusion than donors are. Additionally, thresholds to classify were set at much higher levels (usually around 8 or 10) than our model suggests. However, we can see parallels to other work, where high hydrogen bond basicity was associated with P-gp substrates.[285] Accordingly, Norinder and Haeberlein reported that compounds exhibiting less than five nitrogen and oxygen ([O+N]) entities would readily enter the brain.[286] This threshold corresponds with the cutoff value set in our model.

In the CDK CHAID tree, an increase in rotatable bonds was associated with bad brain permeability. Interestingly, these findings were contrasting to observations of Iyer at al., who stated that increasing molecular flexibility was associated with an increasing permeation. They argued that the number of rotatable bonds were proportional to molecular weigh. Consecutively, an increase in rotatable bonds would clandestinely refer to a relationship of molecular weight and brain permeation. We found a rather low correlation between these two descriptors (R^2 = 0.74) (Figure 26) in our present study and feel that mass, although to a certain extent present, did not contribute substantially to our classification.

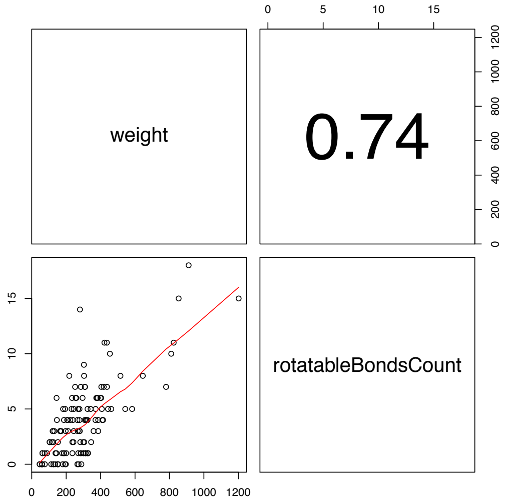


Figure 26 - Scatterplot matrix of intercorrelation molecular weight and rotatable bonds count is shown. The coefficient of determination (R^2) was 0.74.

Remarkably, although the whole feature set was at disposal for training, none of the paradigms selected explicitly molecular weight to classify. This finding is in accordance to the opinion of Abraham and co-workers, who stated a less significant role of the descriptor in predicting brain penetration as certain rules of thumb imply.[287] Diminished permeation capacity with increasing number of rotatable bonds could also refer to potential conformational bulkiness of a molecule. Rotatable bonds are defined as any single bond not involved in a ring structure or connected to a non-terminal heavy atom. In fact, a high number of rotational bonds implied that an extended conformation could roll up into spherical shape. In other words, a molecule could potentially permeate the BBB worse than its molecular weight would indicate owing to a bulky shape. The number of rotatable bonds would then add additional information to models by taking also account of geometrical features rather than simply considering molecular mass. The importance of geometry in predicting brain penetration was substantiated by other DTIs. Our strongest model (CHAID trained on DRAGON descriptors) included edge adjacency matrix indices (SpMAD EA[bo]) to give information on molecular connectivity. The three-dimensional (3D) topological autocorrelation (TDB 05 p, weighted for polarizability) selected by CART represents the information gain from comparing topological and Euclidian distances of atoms in a molecule. Again, this descriptor refers to molecular conformation, as folded chains will have much lower values than stretched ones.[288] Spectral indices like the mass weighted Burden matrix (BCUTS) refer to molecular topology and complexity. Moreover, both DTI paradigms selected 3D-MoRSE descriptors from DRAGON features. These autocorrelation descriptors consider the three-dimensional molecular representation based on electron diffraction patterns. Schuur and co-workers pointed out their use as a measure for mass distribution and branching of a molecule.[289] Although different weighting schemes were used (polarizability for CHAID and intrinsic state for CART), both trees associated lower values with good permeation.

Fragment based approach

The fingerprints selected by binary ACO classification reconfirmed our findings from descriptor based machine learning. The repeated inclusion of ring features indicates a strong contribution of lipophilicity, which is involved in passive and active transport processes across the BBB. Heteroatoms were present in seven out of nine fingerprints, of which four included explicitly nitrogen and/or oxygen atoms. This could relate, in analogy to our findings using DTI, to hydrogen bonding capacity and polarity of a molecule. However an interesting structural feature was fingerprint No. 150, which refers to anticlockwise chirality. To our knowledge stereoselectivity has not yet been used to predict brain penetration capacity. However, *in vivo* studies confirm involvement of stereoselectivity of drug uptake on the BBB.[290, 291]

Conclusions

Our decision trees reconfirmed the involvement of lipophilicity, size, and charge in predicting brain penetration. Additionally, we shed light on features such as molecular geometry, connectivity, and stereochemistry, which are less commonly used in the field.

One could argue that the data underlying our models was derived from rodents and might not accurately reflect the situation in humans. Due to ethical constraints LogPS measurements in man are not feasible. There is little data from intra-operative microdialysis experiments conducted in patients who underwent neurosurgery. Hence, these reports do most likely reflect pathophysiological conditions and are therefore inapt to model the healthy BBB.

Rats are commonly used as an animal model to conduct pharmacokinetic experiments. However, performing *in vivo* perfusion studies is time consuming, costly and requires experience. In addition, these experiments are highly invasive and stand in contrast to general attempts to reduce animal tests. Our models are suitable to predict drug brain permeation in wild type rats and could therefore contribute also to save animal numbers.

Acknowledgements

Claudia Suenderhauf was supported by the Swiss National Foundation (Grant No. 323530-119218).

5.3 Physicochemical and Structural Requirements for Predicting Drug Excretion in Human Breast Milk

Claudia Suenderhauf, Jörg Huwyler, and Felix Hammann

Pharmaceutical Technology
Department of Pharmaceutical Sciences, University of Basel, Klingelbergstrasse 50,
CH-4056 Basel, Switzerland

Submitted to Journal of Clinical Pharmacology & Therapeutics

Abstract

It is commonly agreed that breastfeeding is highly beneficial for mother and child. However, maternal medication intake can impose a risk for the baby as many drugs appear in mother milk. Despite its social, economic, and medical impact, safety of drug intake in nursing is still a relatively unexplored field. We aimed to computationally model drug passage into mother milk. The work presented here is a novel approach to holistically assess active and passive transport processes in the mammary gland epithelium.

We assembled a dataset of small molecule drugs (n=90) and used the DRAGON toolkit and the Chemical Development Kit (CDK) to generate physicochemical descriptors. From these feature sets decision trees were grown. Performance in models using DRAGON features was outstanding and achieved corrected classification rates of 85.3% - 95.3%. Prominent splitting criteria were descriptors of molecular size, branching, charge and geometry. A fragment-based analysis revealed structural elements referring to polarity and to involvement of an active transport component.

We consistently observed strong predictive power in all of our models, which underlines the viability of the present dataset. Descriptor and fragment based models shed light on the molecular requirements to identify safe drugs in nursing. The classifiers presented here will ease decision making in clinical settings or drug design even if no experimental data are available.

Introduction

See Section 3.2.1.3.

Drugs accumulating in mother milk are likely to become increasingly available to the baby. Considering the immature metabolism and drug clearance of the baby, serum levels could rise above therapeutic concentrations and cause side effects. Therefore, models including actively transported structures, which could lead to drug accumulation, are of substantial importance, both in drug development and safety monitoring in the clinical setting.

Materials & Methods

Dataset

We assembled a dataset of 90 compounds for which information on human breast milk excretion profiles was available. The initial dataset was based on 162 marketed small molecule drugs from the therapeutic drug database.[230] The source supplied either qualitative (e.g., descriptive) or quantitative (numeric data) information on endpoints. We therefore classified compounds accordingly to Table 23 as excreted (BM^+, n = 40) or non-excreted (BM^-, n= 50) drugs.

Class label	Qualitative label from source	Quantitative label from source	n
BM^+	Accumulation	Milk levels > maternal blood levels	40
	Same/ greater/ higher than maternal blood levels	Milk levels 1-40x higher than maternal blood level	
	Therapeutic concentrations, Freely diffusing	Milk/maternal blood ratio 1.5 or 8.5:1	
	Clinically significant/ extensively secreted		
BM^-	Traces, minimally, negligible, not secreted	<= 1% of maternal blood level	50
	Very/ extremely low/ small quantities		

Table 23 - Class label assignment was performed as shown. The source supplied information on drug presence in mother milk either qualitatively (e.g., nominal quantification) or quantitatively (e.g., milk/plasma ratios). Due to varying detection techniques, maternal drug concentrations were measured either in plasma, serum or blood. For reasons of readability, we refer to these different values as maternal blood levels. We denoted compounds passing into breast milk as BM+ (n = 40) and non-permeating molecules as BM- (n=50).

Compounds not meeting these requirements were exempt from learning, for the sake of classifier quality. We did not exclude any compounds which are known substrates of active transport and/or are known to accumulate in milk.

The data source contained generic names. Corresponding structures were retrieved from the National Library of Health database, PubChem.[292] As the paradigms used cannot handle disconnected structures, we removed counterions form salts prior to their consideration.

Descriptors

See Section 4.2 and 4.6.3.

To derive descriptors from chemical structure we used the commercial DRAGON toolkit. To ease reproducibility of our models we also applied descriptors calculated with the open source CDK software.

Chemical fingerprints

See Section 4.1.3.

To get insight into structural requirements for safe drugs in nursing, we performed a fragment-based analysis.

Decision tree induction

See Section 4.3.1.

In this study, we set maximum tree depth to five nodes for CART, and to three nodes for CHAID, respectively. We allowed a minimum of 10 instances in the parent and 5 cases in the child node.

Ant colony optimization

See Section 5.5.

Chemical Space

We assessed the chemical space spanned by our dataset calculating the Tanimoto coefficient spanned by the MACCS keys using the Open Babel toolkit and in-house software.[293] Keys of every molecule were compared to the remaining set. The resulting similarities were then averaged over the whole dataset (n=90) to receive an overall similarity value.

Quality measure

See Section 4.5.

Cross-validation

See Section 4.6.2.

Software used

See Section 4.6.3.

In-house software was available for ACO (Section 5.5.).

Results

Chemical Space

We calculated an overall similarity of 0.333 for our dataset using Tanimoto coefficient. This is a low level of similarity and represents a reasonable chemical space on which to base machine learning. Models created will most likely interpolate well for future small molecule drugs.

Performance of machine learning methods

Both DTI paradigms achieved the strongest models when trained on DRAGON descriptors. Classification and regression tree (CART) produced the best model in our study, achieving a CCR of 95.3% (MCC: 91%) closely followed by Chi-squared interaction detector (CHAID), which performed with a CCR of 92.5% (MCC: 87.1%). Figure 27 and Figure 28 show the corresponding trees.

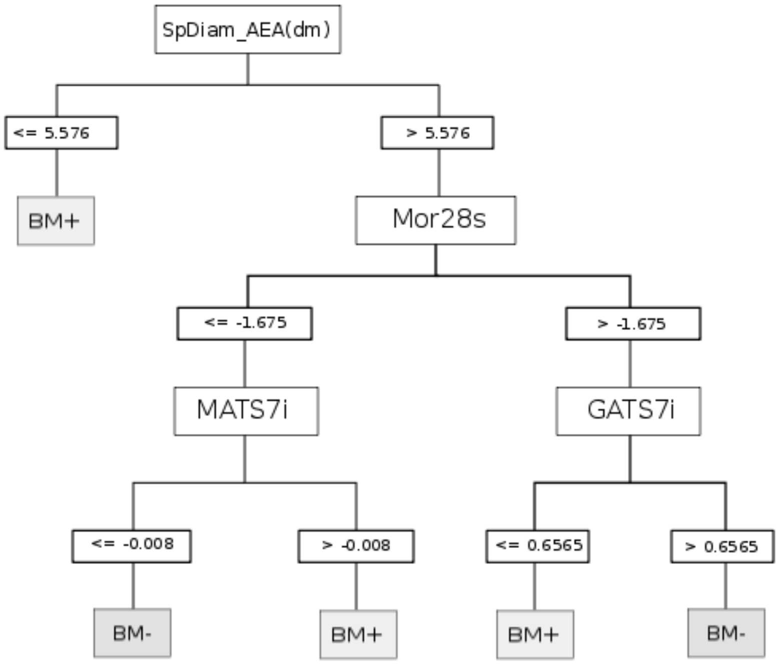


Figure 27 - We show the CART built on DRAGON features. The tree predicted with CCR of 95.3% (MCC: 91%). SpDIAM_AEA(dm) is the spectral diameter from augmented edge adjacency (AEA) matrix, weighted by dipole moment. Mor28s stands for the MoRSE3D descriptors at lag 28, weighted by I-state. MATS7i and GATS7i are 2D autocorrelation matrices at lag 7 weighted by their ionization potential (i), defined by Moran and Geary, respectively. BM^+ and BM^- denote permeating and not permeating compounds, respectively.

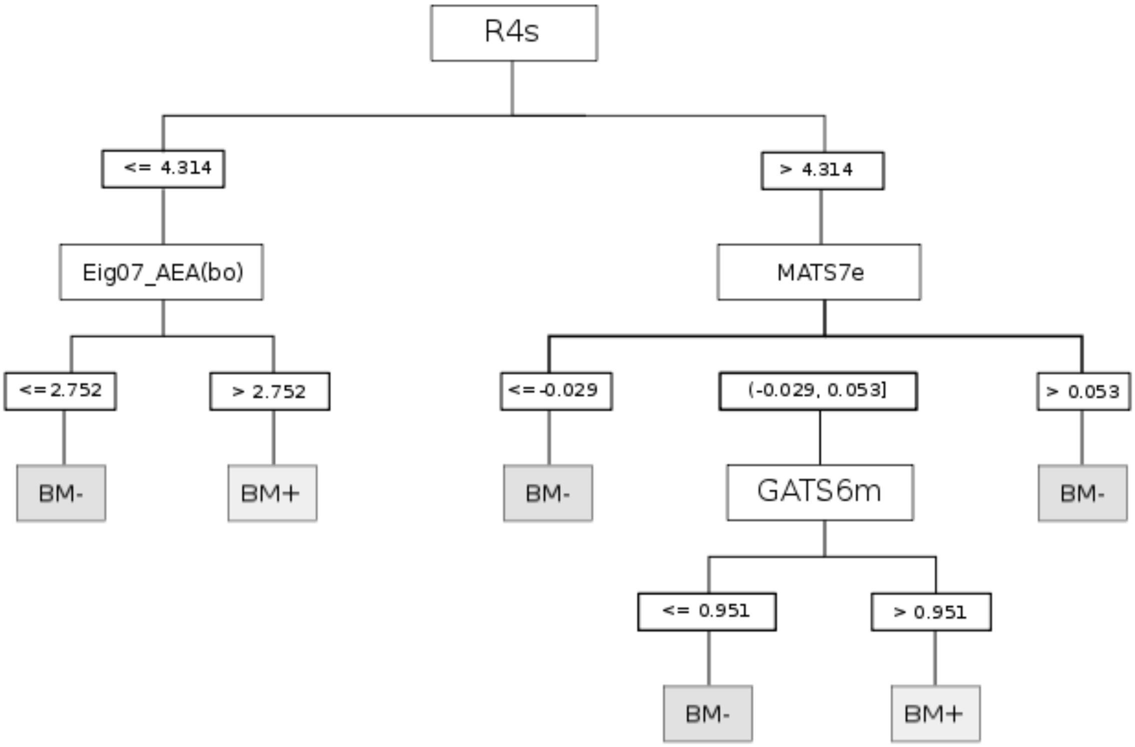


Figure 28 - The CHAID tree built on DRAGON descriptors is shown. The paradigm achieved a performance of CCR: 92.5% (MCC: 87.1%). First splitting criterion was R4s, the R autocorrelation at lag 4 weighted by intrinsic state from the GETAWAY descriptors. Eig07_AEA(bo) is a descriptor of the edge adjacency indices. It stands for the eigenvalue number 7 from the augmented edge adjacency matrix weighted by bond order. MATS7e and GATS6m are both 2D autocorrelations, defined by Moran and Geary. MATS was weighted for electronegativity and GATS was weighted for mass. BM^+ indicates permeating and BM^- not permeating compounds.

Performance in these models was outstanding, but to a certain extent at expense of interpretability. When we trained the paradigms on CDK features, more intuitive features were chosen.

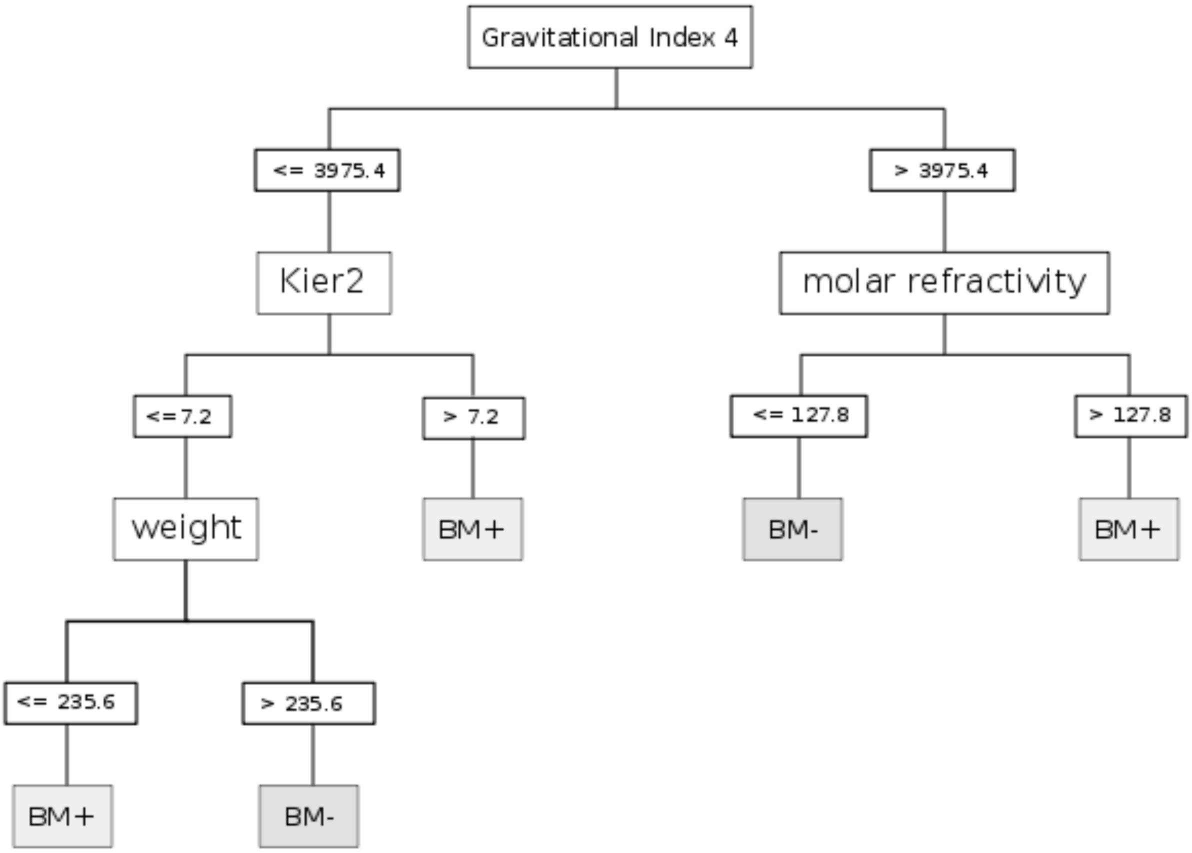


Figure 29 - When CART performed on CDK features the depicted tree was grown. Splitting criteria were gravitational Index 4, Kier Hall kappa shape index, molecular weight, and molar refractivity. The paradigm achieved a performance of CCR: 85.3% (MCC: 70.2%). BM$^+$ stands for permeating molecules and BM$^-$ for not permeating ones.

However, predictions of CART performed with CCR of 85.3% (MCC: 70.2%)(Figure 29) and CHAID with CCR of 85% (MCC: 70.7%)(Figure 30). Table 24 summarizes DTI performance on all feature sets.

Descriptors	Paradigm	CCR(%)	MCC(%)	SENS	SPEC
DRAGON	CHAID	92.5	87.1	1	85
	CART	**95.3**	**91**	**98**	**92.5**
CDK	CHAID	85	70.7	80	90
	CART	85.3	70.2	92.5	78

Table 24 - We present the performance of DTI on DRAGON and CDK descriptors. Corrected classification rate (CCR) and Matthews correlation coefficient (MCC) for all DTI models are given.

CART tree trained on DRAGON features achieved the best performance (indicated in bolt face). Sensitivity and specificity are given as SENS and SPEC, respectively.

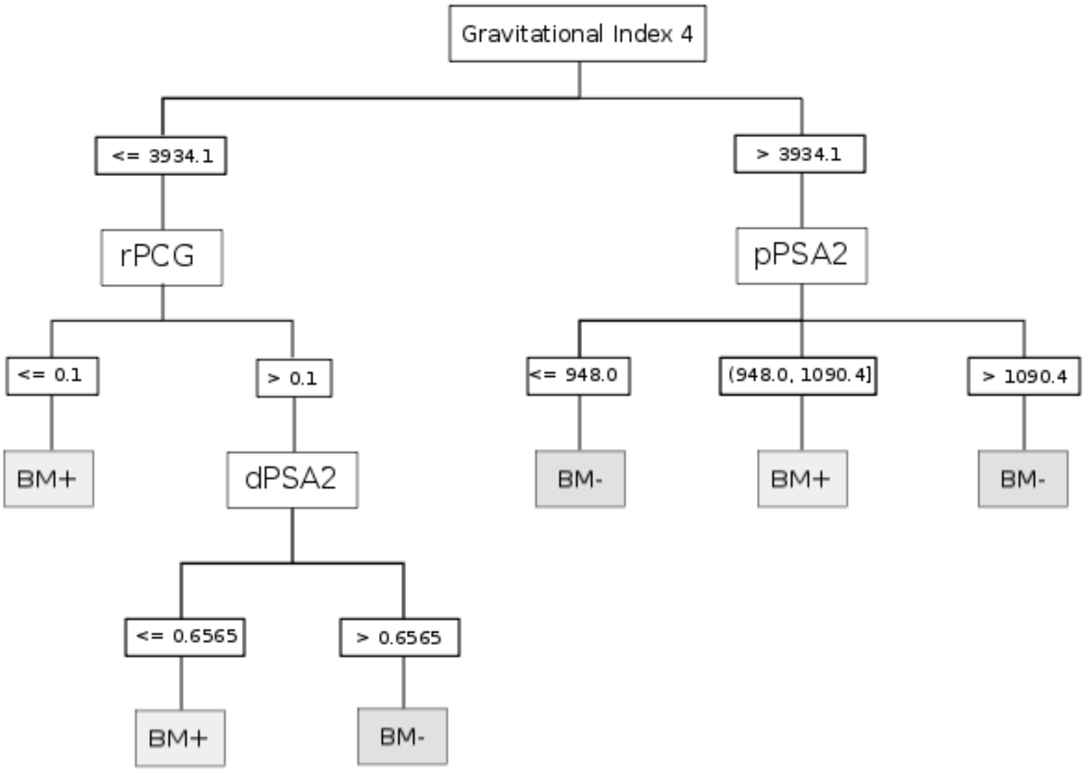


Figure 30 - We present the tree grown with CHAID on CDK descriptors. The paradigm selected gravitational index four, relative positive charge (rPCG), partial positive surface area multiplied by total positive charge on the molecule (pPSA2), and the difference of pPSA2 divided by molecular surface and partial negative surface area multiplied by total negative charge on the molecule (dPSA2). RPCG, pPSA2, and dPSA2 belong to the charged polar surface areas descriptors (cPSA). CHAID achieved a performance of CCR: 85% (MCC: 70.7%). BM+ indicates permeating and BM- not permeating drugs.

Our analysis of structural requirements for safe drugs in nursing revealed 10 relevant fingerprints which are listed in Table 25. The molecular fragments performed with a CCR of 83% (MCC: 66%) and an area under the curve (AUC) of 0.85. The corresponding ROC curve is depicted in Figure 31.

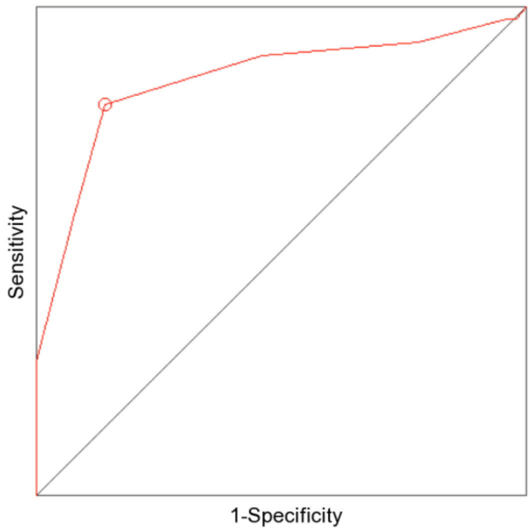


Figure 31 - The ROC curve achieved of best-performing ACO model is shown. The 10 fingerprints from the MACCS keys performed with a CCR of 83% (MCC: 66.2%). The area under the curve (AUC) was 0.85. The cutoff point taken as maximum Youden's J is indicated by a circle.

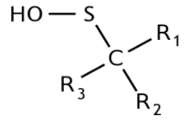
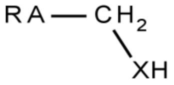
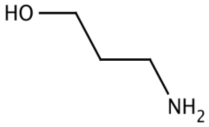
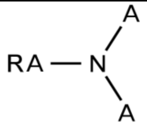
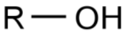
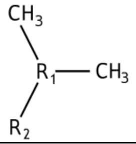
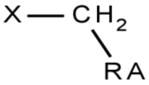
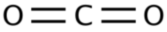
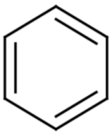
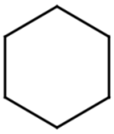
No	Sample Structure	SMARTS	Description
51		[#6]~[#16]~[#8]	Carbon bound sulfoxide.
82		*~[CH2]~[!#6;!#1;!H0	Methylene connected to a heteroatom with at least one hydrogen atom and to an additional molecule.
97		[#7]~*~*~*~[#8]	Nitrogen connected to oxygen via any three atoms.
122		*~[#7](~*)~*	Any atom connected to nitrogen. Nitrogen has to be connected with any two additional atoms.
139		[O;!H0]	Oxygen with at least one hydrogen (e.g., hydroxy group).
149		[C;H3,H4]	More than one methyl group.
153		[!#6;!#1]~[CH2]~*	Heteroatom with methylene group, connected to a rest.
159		[#8]	More than one oxygen atom.
162		a	Aromatic atom.
163		*1~*~*~*~*~*~1	Six-membered ring structure, occurring twice in molecule.

Table 25 - The 10 best performing MACCS Keys selected by ACO are shown. We give the number (No) in the fingerprint set, a sample structure, the corresponding SMART keys, and a short explanation of the chemical fragment.

Discussion

Our dataset exhibited a reasonable dissimilarity to interpolate for existing and future small molecule drugs. This is an advantage, if safety estimates are needed and no experimental data are available. In terms of size, our dataset might be considered modest compared to others. However, division in test and training sets dramatically reduces the effective size of data used for creating models. We overcame this problem by cross-validating our models, thus using the whole chemical space available. Additionally, k-fold cross-validation resulted in a realistic performance estimate as the data subsets were randomly composed. Therefore, we considered our models in terms of size competitive with other models using holdout validation.

We felt that a numeric prediction approach would not be sensible due to the inherent fuzziness in reporting quantities of compounds detected in mother milk. We therefore decided to restore discriminatory power by partitioning compounds into two classes.

Classifiers trained with DRAGON descriptors achieved excellent performance. However one could argue that the selected splitting criteria are not easily interpretable and traceability of trees is hampered to a certain extent. Although the CDK trees could not hold with the strong performance given in DRAGON based models, they yielded more intuitive features. Interestingly, these splitting criteria did largely reflect our current knowledge of the process.

Typically, molecular size and weight substantially determine the membrane permeation capacity. While small compounds readily undergo passive diffusion, bulky ones will be sterically hindered. An earlier study by Meskin and co-workers revealed a negative correlation of molecular weight and milk plasma ratios.[294] We could reconfirm his observation in our CART model trained with CDK descriptors, where lower values were associated with well-permeating compounds.

Additionally, CART classified drugs exhibiting molar refractivity over the critical threshold as permeating drugs (BM+). Molar refractivity refers to molecular shape and compactness. Tightly packed compounds would more readily permeate through membranes than highly branching ones. Gravitational index 4 considers the mass distribution related to intramolecular distances accounting for the bulk cohesiveness. Information on molecular graph complexity gave topological shape descriptors of second order (Kappa shape Index, Kier 2) included by CART. Other groups confirmed the role of molecular density and complexity in the permeation capacity.[178, 295]

Lipophilicity plays a delicate role in determining permeability. It is certainly a prerequisite for membrane interactions and permeation into milk. However, in maternal circulation, high lipophilicity hampers solubility and enhances serum protein binding, counteracting a compound's distribution and eligibility to be secreted. None of our models selected partition coefficient (LogP), although some seminal work accounted for the descriptors dominant role in estimating drug diffusion into breast milk.[63, 296] In these studies, the number of features was restricted by an expert-based preselection. We agree with

Agatonovic and co-workers, who stated that with higher dimensional descriptors at hand, machine learning paradigms can select more refined measures of lipophilicity than LogP.[295]

Molecular polarity played a key role in models trained with CDK descriptors. Polarity refers to a compounds hydrophilic potential and can be seen as an indirect measure for lipophilicity. Namely, charged polar surface area descriptors were prominent splitting criteria in the CHAID tree. We observed a certain tendency in compounds with low polarity (i.e., more lipophilic drugs) to better permeate into mother milk. However, thresholds set in our models were subtle and reflected the narrow scope of this attribute.

All splitting criteria selected from the DRAGON set relied on weighted matrices. Weighting schemes are used to encode chemical information on bonds or molecules that is not contained in a molecular graph. SpDiam_AEA(dm) and Eig07_AEA(bo) belong to the edge adjacency indices, where spectral diameter and eigenvalue of the augmented edge adjacency (AEA) matrix are considered. Bond matrices based on the chemical graph theory encode information on intramolecular connectivity. The weighting schemes considered bond order (bo) and dipole moment (dm). The importance of connectivity and molecular polarity corresponds with our findings from CDK descriptors.

Spatial autocorrelation descriptors, such as Moran's (MATS) and Geary’s (GATS) autocorrelation were chosen by both paradigms. Autocorrelation descriptors evaluate a certain atomic property for every atom in the molecule. Both features selected by CART used molecular ionization potential (i) as weighting property. The involvement of ionization could refer to an ion-trapping effect as breast milk exhibits a lower pH than human plasma. Electronegativity (e) and molecular mass (m) were the weighting schemes selected by CHAID. Observations from CDK learned trees confirmed involvement of charge and mass in predicting BM-/BM+ compounds. Two features were weighted by intrinsic state. Namely, these were R4s from the GETAWAY descriptor containing information on molecular branching and Mor28s from the 3D MoRSE descriptors which encodes three-dimensional information based on electron diffraction.[297, 298] Intrinsic state gives information on the electrotopological state of a molecule.

Our fragment-based approach gave information on structural requirements for safe drugs, e.g., poorly permeating compounds. The repeated selection of amides, oxygen, and hydroxy groups indicates involvement of hydrogen bonding capacity and molecular polarity. Other strongly polarizing groups, such as sulfoxyde are likely to oxydize amino acids, e.g., cysteine, at the binding site of transporters through irreversible addition of a thiol group.

Interestingly, fingerprint fragments revealed by ACO included six-membered rings and aromaticity. We found these features also in typical substrates of BCRP such as chemotherapeutics (e.g., mitoxantrone), antivirals (e.g., zidovudine), and antibiotics (e.g., ciprofloxacin).[299] However, there is evidence that the mammary gland epithelium expresses a multitude of transport proteins including other ABC transporters such as P-gp (MDR1).[300] In contrast to BCRP, studies indicate an apical to basolateral transport, which would counteract drug accumulation in breast milk.[104] As both proteins share certain substrates, the complexity of molecular interplay may be potentiated. In earlier models of metabolic enzymes and P-gp, constitutional and connectivity measures played key roles to predict substrates.[177] Thus, we were not surprised to find parallels between our results and those from other models of these transporters.

Conclusions

Excretion of drugs into breast milk remains poorly understood. Beside passive diffusion form plasma to milk, other transport mechanisms exist for various compounds, as evidenced by the emergence of further transport proteins that play a role in mother milk composition. Excretion and reabsorption processes additionally confound the issue. Therefore, our approach to predict breast milk excretion beyond separation into active and passive processes seems sensible. Although our models do not compensate for careful assessment of single biological mechanisms involved, they are able to predict the drug presence in human breast milk.

Acknowledgements

Claudia Suenderhauf was supported by the Swiss National Foundation (Grant No. 323530-119218).

5.4 A Computational Assessment of MRP2: Prediction of Substrates, Inducers, and Inhibitors.

Claudia Suenderhauf, Felix Hammann, and Jörg Huwyler

Pharmaceutical Technology
Department of Pharmaceutical Sciences, University of Basel, Klingelbergstrasse 50,
CH-4056 Basel, Switzerland

Submitted to Journal of Drug Metabolism and Pharmacokinetics

Abstract

Multidrug resistance protein 2 (MRP2) plays an important role in drug clearance and efflux. As both, the functional loss and overexpression were associated with pharmacokinetics alteration of several drugs, knowledge on structural and physicochemical requirements for MRP2 drug interaction is desirable. We performed a computational analysis based on a small molecule datasets from literature providing information on MRP2 inhibitors (n = 277), inducers (n= 122) and substrates (n= 76).

Decision trees were induced and resulting models predicted inhibitors and inducers with a corrected classification rate (CCR) of 87.3 - 87.6%, 84 - 90.9%, respectively. Our tree models used descriptors of charge and molecular complexity along with connectivity measures to predict MRP2 inhibitors and inducers.

Introduction

See Section 3.2.3.1.

Despite the clinical implications of MRP2 on drug excretion, knowledge on requirements for MRP2 interaction is mostly restricted to results of single compound analysis. Quantitative structure activity relationship (QSAR) analysis is ideally suited to summarize these data to get more generalized information on physicochemical properties involved. Moreover these techniques enable implementation of fast and accurate screening techniques for future drug development. Computational methods have already been successfully applied to model drug behavior of other ABC transporters, such as P-gp and BCRP.[258, 299, 301] It was our intention to contribute to the ongoing discussion on the chemical requirement for MRP2 interaction, by presenting a decision tree approach to predict this highly complex endpoint. The production of easily interpretable rules is certainly an advantage of decision tree induction. They can easily be implemented to apply them on unseen compounds to predict MRP2 inhibitors, inducers and substrates.

Materials & Methods

Data sets

We assembled three datasets of small molecules from published literature. The dataset concerning MRP2 inducing compounds consisted of 122 entities. We found 52 inducers and labeled them as “MRP_{ind}+”. All molecules that were explicitly found to not induce MRP2 expression received the labeled “MRP_{ind}-”.

We collected a set of 277 compounds where information of their inhibitory function on MRP2 was available. We found 146 Inhibitors (label = “$MRP2_{inh}$+”) and 131 non-inhibitors (label= “$MRP2_{inh}$-”). The data set for MRP2 substrates consisted of 77 molecules, of which only 10 were explicit non-substrates.

Descriptors

See Section 4.2.

Decision Tree Induction

See Section 4.3.1

We used classification and regression trees (CART) to induce trees for the current study.[204] A prior feature reduction was not needed as decision trees perform implicitly by themselves a feature selection. To grow trees we restricted splitting in parent nodes to a minimum of five instances and in child nodes to a minimum of two cases. CART was grown to a maximum depth of five levels.

Quality measures

See Section 4.5.

Cross-validation

See Section 4.6.2.

Software

See Section 4.6.3.

Results

MRP2 Inducers:

The best performing trees for MRP2 inducers were created on DRAGON features (CCR: 90.9%). The paradigm produced models with 83.5% CCR on CDK features. Figure 32 shows the corresponding tree.

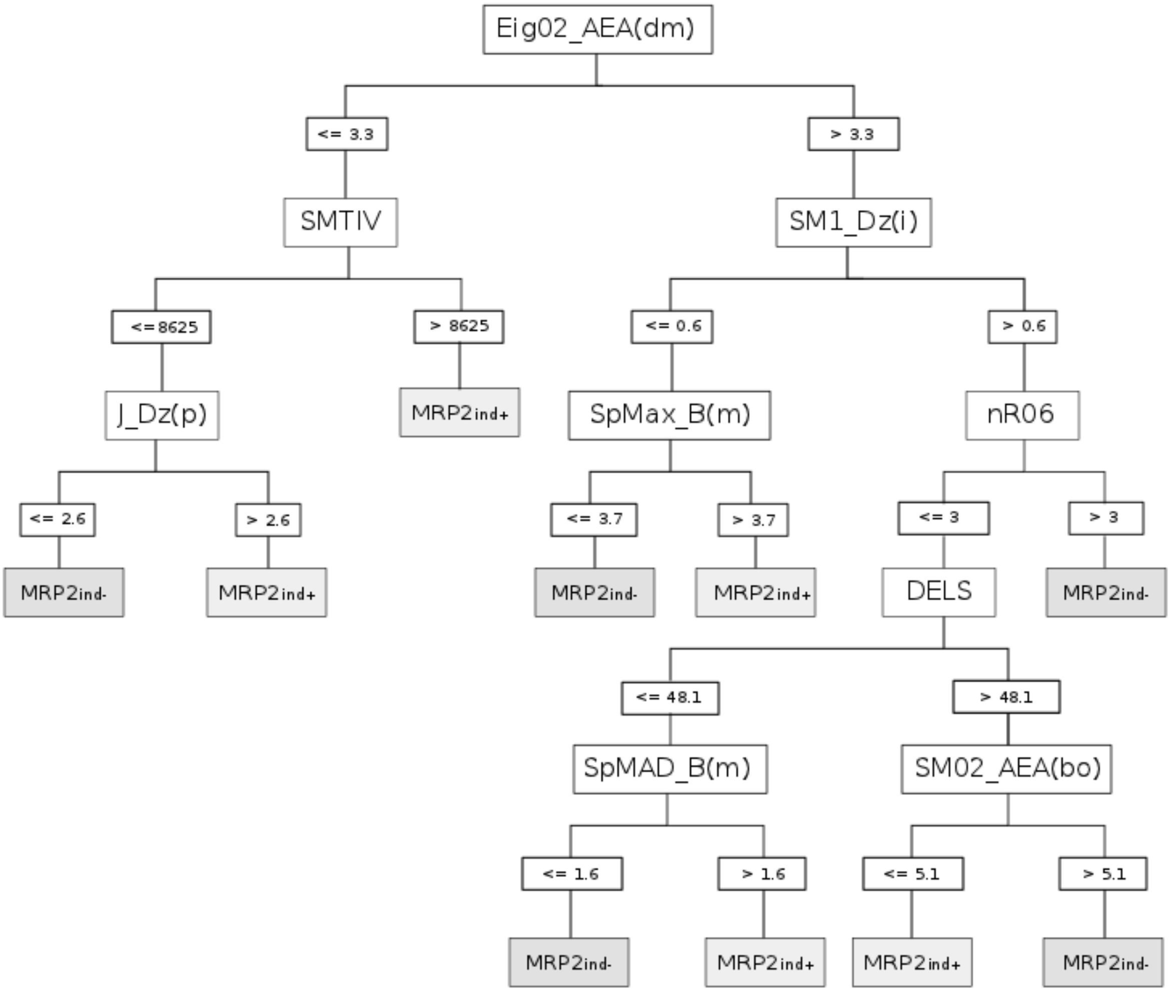


Figure 32 - Classification and regression tree (CART) trained on DRAGON features. The paradigm achieved a corrected classification rate (CCR) of 90.9%. Splitting criteria were eigenvalue of the augmented adjacency matrix weighted for dipole moment (Eig02_AEA[dm]), Schultz molecular topological index by valence vertex degrees (SMTIV), Balaban like index from the Barysz matrix weighted by polarizability (J_Dz[p]), spectral moment from the Barysz matrix weighted by ionization(SM1_Dz[i]), mass weighted largest Burden eigenvalues (SpMax_B[m]), number of six membered rings (nR06), molecular electrotopological variation (DELS), mass weighted Burden matrix (SpMAD_B[m]), bond order weighted spectral moment of the augmented adjacency matrix (SM02_AEA[bo]). MRP2ind+ indicates MRP2 inducers while MRP2ind- indicates non-inducers.

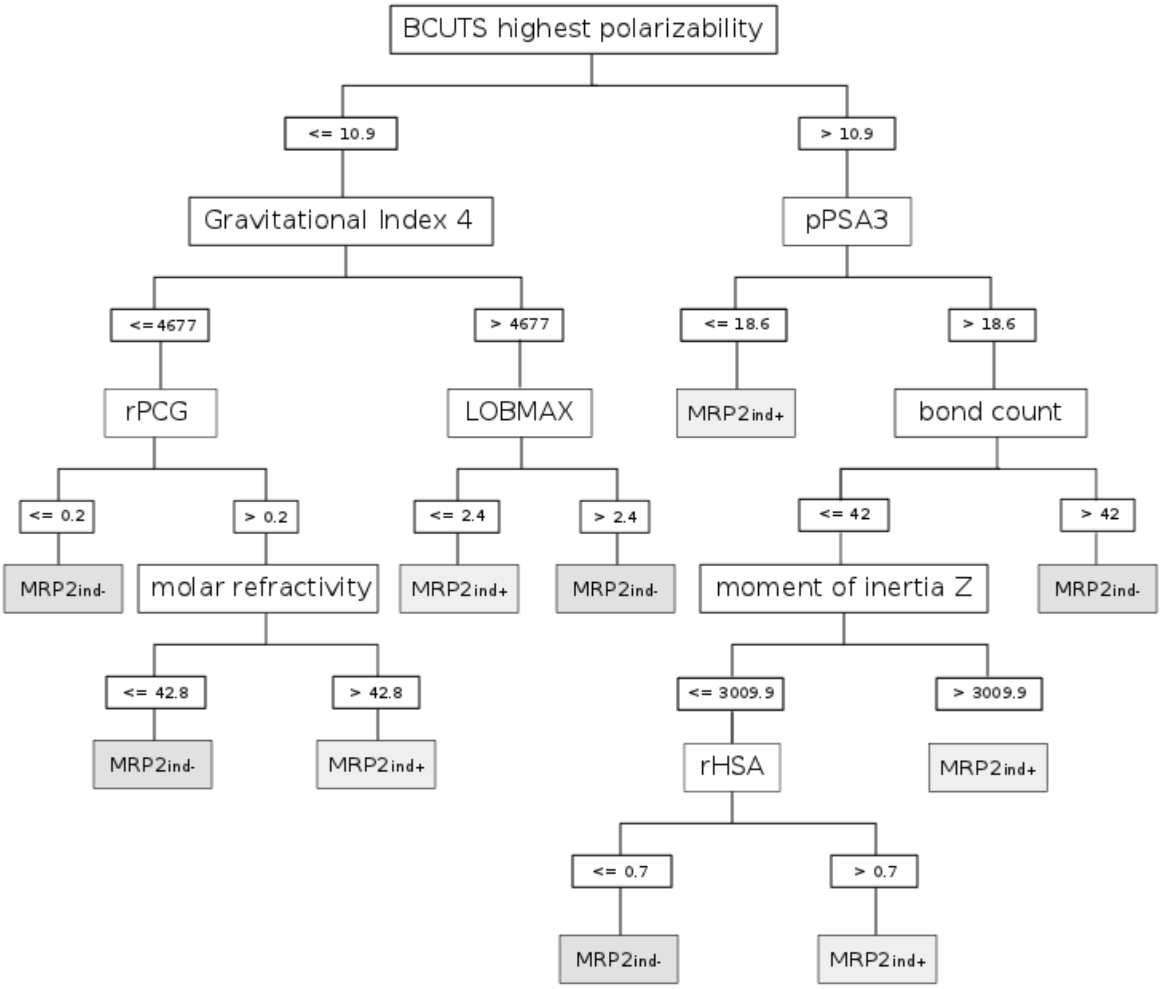


Figure 33 - Decision tree induced on CDK features is shown. The tree achieved a correct classification rate of 83.5%. Features selected were polarizability weighted BCUTS, gravitational Index four, relative positive charge (rPCG), molar refractivity, charge weighted partial positive surface area (pPSA3), bond count, moment of inertia on the z-axis, and relative sum of solvent accessible surface areas of atoms with absolute value of partial charges less than 0.2 (rHSA). MRP2ind+ and MRP2ind- indicates MRP2 inducers and non-inducers, respectively.

MRP2 Inhibitors:

The best model for MRP2 inhibitors was produced with the CART paradigm on DRAGON features, yielding a CCR of 87.6% (Figure 34). The corresponding tree using CDK features predicted the endpoint with 87.3% CCR (Figure 35).

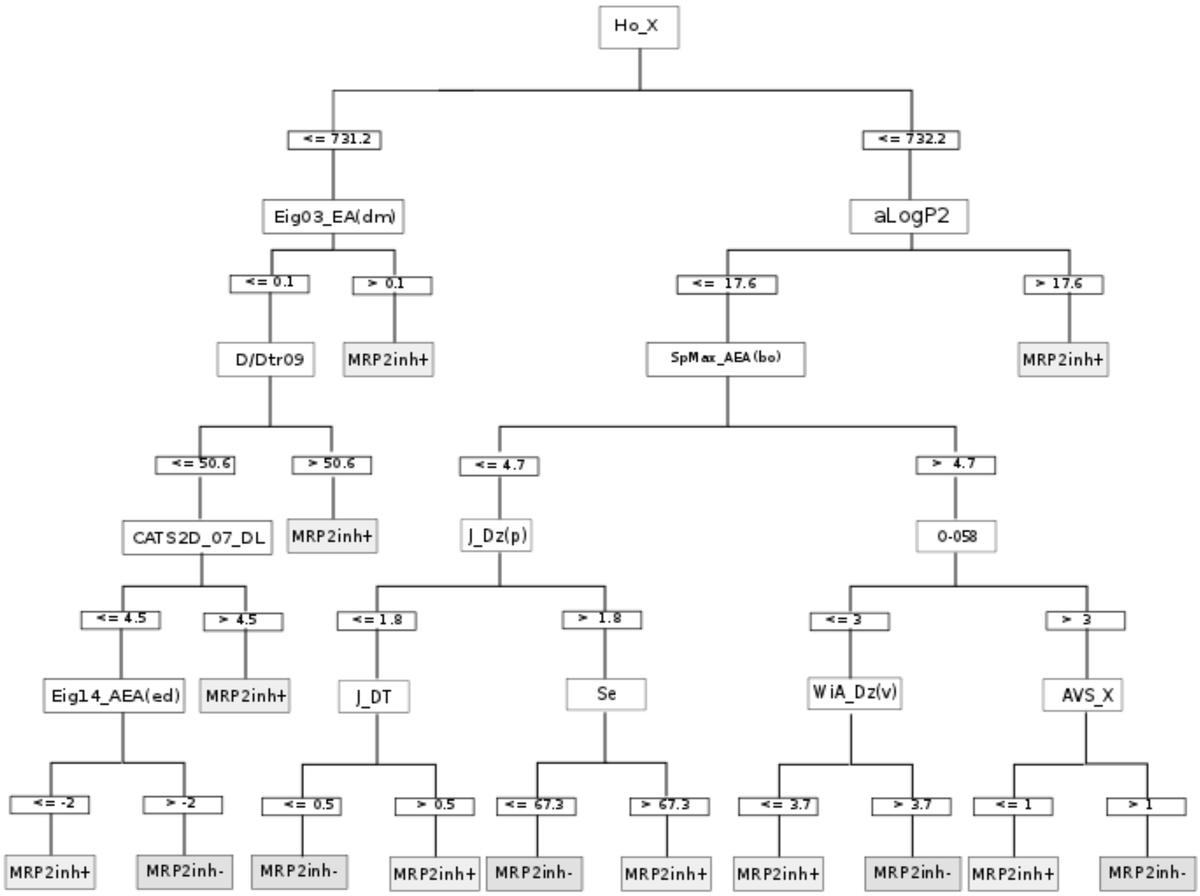


Figure 34 - The best decision tree predicting MRP2 inhibitors on DRAGON features with corrected classification rate (CCR) of 87.6% is shown. Splitting criteria were Hosoya like index form the chi matrix (Ho_X), dipole moment weighted eigenvalue of the edge adjacency matrix (Eig03_EA[dm]), distance/detour ring index of order 9 (D/Dtr09), pharmacophore pair distance of hydrogen donors and lipophilic groups (CATS2D_07_DL), edge degree weighted eigenvalue of the augmented adjacency matrix (Eig14_AEA[ed]), squared Moriguchi octanol water partitioning coefficient (sLogP2), bond order weighted leading eigenvalue of the augmented adjacency matrix (SpMax_AEA[bo]), polarizability weighted Balaban like index from the Barysz matrix (J_Dz[p]), sum of Sanderson electronegativity (Se), double bound oxygen count (O-058), average van der Waals volume weighted Wiener-like index from the Barysz matrix (WiA_Dz[v]), average vertex sum of the chi-matrix (AVS_X). MRP2inh- and MRP2inh+ indicates non-inhibitor and inhibitor, respectively.

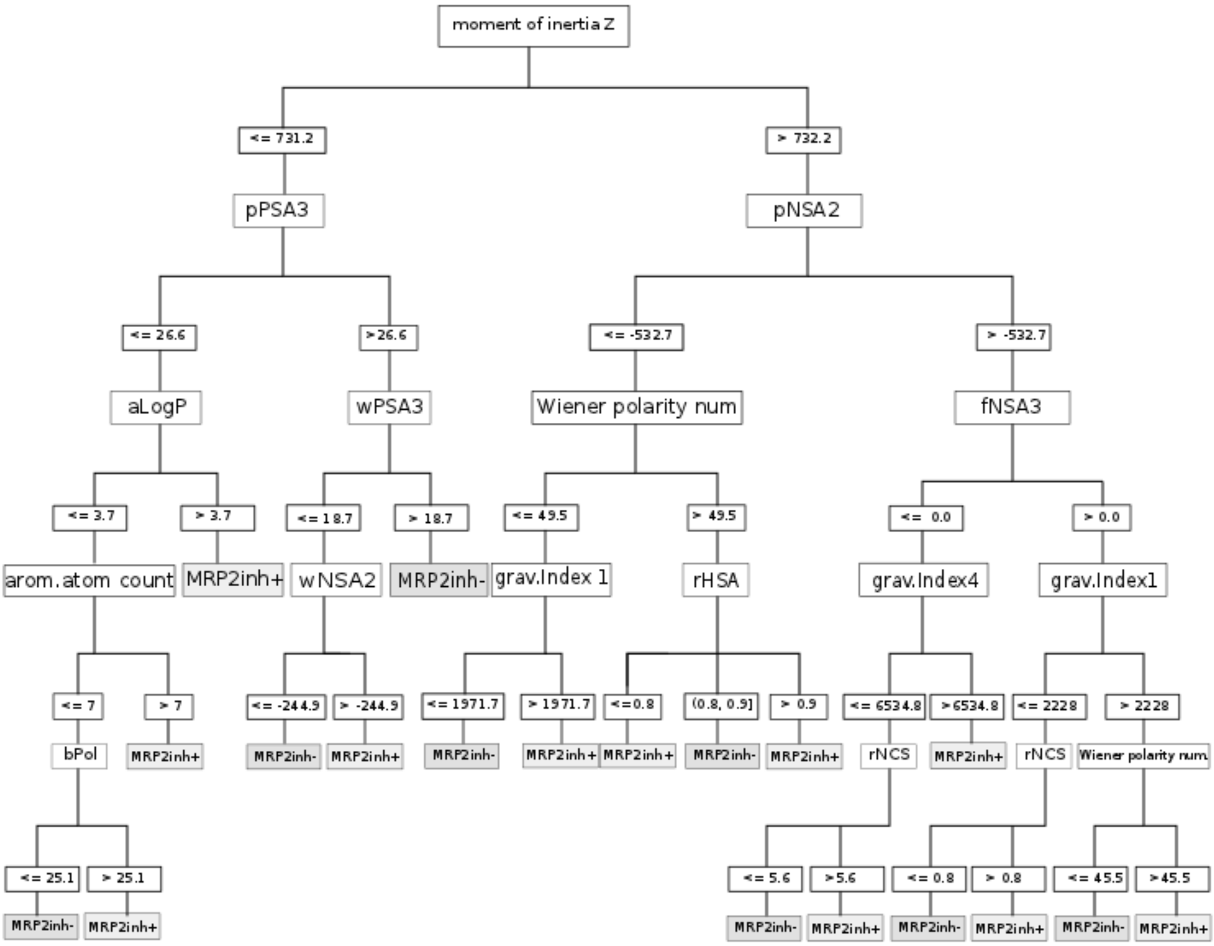


Figure 35 - Classification and regression tree trained on CDK features is shown. MRP2 inhibitors and non-inhibitors are indicated as MRP2inh+ and MRP2inh-, respectively. Features selected include moment of inertia in the z-axis, charge weighted partial positive surface area (pPSA3), partition coefficient (aLogP), aromatic atom count, the sum of the absolute difference between atomic polarizabilities of all bonded atoms in the molecule (bPol), charge weighted partial positive surface area (wPSA3) and partial negative surface area (wNSA2) multiplied by molecular surface, partial negative surface area multiplied by total negative charge (pNSA2), Wiener polarity number (Wiener polarity num.), gravitational index one and four (grav. Index 1, grav. Index 4), relative sum of solvent accessible surface areas of atoms with absolute value of partial charges less than 0.2 (rHSA), charge weighted partial negative surface area divided by total surface (fNSA3), and relative negative charge (rNCS). The paradigm performed with a corrected classification rate (CCR) of 87.3%.

Performance of CART predicting MRP2 inhibitors and inducers on both feature sets is summarized in Table 26.

Dataset	Feature Set	CCR (%)	MCC (%)	SENS (%)	SPEC (%)
MRP2 Inducers	CDK	83.5	66.8	94.2	72.9
	DRAGON	**90.9**	**83.4**	**84.6**	**97.1**
MRP2 Inhibitors	CDK	87.3	76.0	95.2	79.4
	DRAGON	**87.6**	**76.8**	**95.9**	**79.4**

Table 26 - A summary of the 10-fold cross-validates decision trees on MRP2 inducers and inhibitors is given. Trees were trained on DRAGON and CDK features. Corrected classification rate (CCR), Matthews correlation coefficient (MCC), sensitivity (Sens), and specificity (Spec) are indicated. Best models are highlighted in bold letters.

MRP2 Substrates:

Our dataset of MRP2 substrates was strongly skewed, in a way that only 10 compounds were judged to be non-substrates. The number of these negative compounds would be by far to small to make adequate assumptions on their chemical requirements. Although our decision trees could discriminate between these unbalanced groups, interpretability was severely reduced because of the small class size (data not shown).

Discussion

We presented three new datasets assembled form literature for MRP2 inhibitors, inducers, and substrates. The data consisted only of experiments conducted on human MRP2. Compared to other transporters, reports on MRP2 drug interaction were relatively rare in literature. Ideally, one would create a database on a high-throughput screening. However, these are still missing for MRP2. This might be due to the highly interactive character of transporter and enzyme interplay on level of the polarized cell. The establishment of a potent in vitro essay, such as MRP2 expressing polarized cells, is a challenging endeavor. Interestingly, relevant directed transport occurs only in presence of coexpressed basolateral uptake transport.[302, 303] Moreover, there is evidence that minimal mutations in MRP2 lead to broader substrate specificity and activity.[131, 304, 305] Further analysis of single nucleotide polymorphisms will hopefully lead to more specific essays for substrate testing. Another issue, which is hard to address is the saturation capacity of the transporter. The experimental data published is mostly restricted to a careful analysis of single compounds, which holds that information. However, for the present study, we decided to use a simplified view by binning compounds in positive or negative classes, regarding their activity. This procedure would allow for models with higher discriminatory power, as it would compensate for differences in experimental settings and show a reduced vulnerability for overfitting.

For inhibitors and inducers we assembled well balanced datasets, in terms of positive compounds not outweighing positive ones. They readily qualified for computational analysis. This did not hold for our MRP2 substrate dataset. Although we assembled 66 substrate compounds, only 10 molecules were reported to be explicitly not substrates of MRP2. We feel that non- transported compounds suffer from a positive publication bias, despite their impact for drug development. Although our decision trees could discriminate between these unbalanced groups, interpretability was severely reduced (data not shown). The computational analysis led most likely to a projection of the dataset instead of the effective requirements for MRP2 substrates.

As discussed before, adequate essays for MRP2 substrates involve a multitude of factors such as influx transporters and probably also metabolic enzymes, which transform parent compounds. The proximity and interaction with phase I and II enzymes underlines this assumption. Structures with high lipophilicity and charge are typically metabolized by cytochromes.[177] Such compounds could likely be prone to prior metabolism before undergoing secretion by MRP2. Hence, it would be tempting to assess every metabolism and transport step by a single model and then concatenate these to predict general drug excretion. We feel that such an approach would introduce error accumulation, as every single model suffers from an individual degree of uncertainty. Several studies showed that for creation of robust and noise resistant computational models, simplification to end- and starting point is a valid and beneficial procedure.[259, 260]

Decision trees of MRP2 inducers and inhibitors were highly accurate in discriminating active and inactive compounds. The best model for inhibitors was achieved on DRAGON features, yielding predictions of 87.6% accuracy (Table 26). Inducers were predicted with 90.9% accuracy. This is a remarkable performance as all models were cross-validated. Our models for inhibitors outperformed recent modeling attempts by Zhang et al. and Pedersen et al., who reported an accuracy of 77% and 72%, respectively.[306, 307] In both studies, no cross-validation was applied.

To our knowledge, MRP2 shares many inhibitors and inducers with other ABC transporters like P-gp or BCRP. Certain resemblance of the splitting criteria revealed by our trees with earlier studies concerning P-gp modeling does therefore not surprise. A computational analysis of P-gp interactions revealed an involvement of lipophilicity and aromaticity to discriminate inhibitors form non-inhibitors.[258] A study of BCRP inhibition used exclusively lipophilicity and polarizability to discriminate inhibitory compounds.[308] We are therefore not surprised to find these features in both trees for MRP2 inhibitors. While in models of P-gp inhibitors ring bonds played a dominant role,[258] MRP2 inhibitors seemed to be more dependent on charge distribution, reflected by repeated selection of cPSA descriptors. Interestingly, our findings reconfirmed a recent study of Pedersen at al., where they state a correlation of lipophilicity, polarity, and aromaticity with MRP2 inhibitors. [307] A relatively unusual feature to classify inhibitors on, is the molecular mass distribution, quantified by gravitational indices. Trees trained on CDK features selected the first and the fourth index for discrimination. These indices could be seen as a refined measure for molecular mass introduction. Additionally, the descriptor holds information on molecular connectedness and geometry. Although not commonly used to study ABC transporters, gravitational indices and moment of inertia have been used to model cytochrome P450 (CYP) inhibitors.[177] This could be an indication of the transporter's interplay with phase one enzymes like CYPs. However, we found also a links to hepatotoxicity prediction studies. Drugs associated with toxicity were generally more lipophilic and charged than safe compounds. [259]

MRP2 inducers were generally compacter and bulkier molecules than non-inducers were and showed a tendency for higher polarity. However, we could not generally associate smaller size with inducer activity as proposed by Pedersen.[307] Our CDK features tree indicated a rather fine discrimination, where additional features such as charge distribution and molecular interconnectedness are taken into account. In the tree built with DRAGON descriptors, we saw a tendency of inducers exhibiting lower numbers of six-ring structures, which indicated that they are less lipophilic than non-inducers. We stress that the splitting criterion cannot be interpreted in isolation, as charge weighted connectivity is a dominant factor in this tree.

Interestingly, we found in consistency with a comparable study of P-gp,[258] that decision trees for inhibitors were more complex and deeper than those for inducers. This could reflect the multitude of possible drug interactions leading to transporter inhibition. While it was argued that MRP2 inducers would bind on one of the transporter’s two binding sites, the example of P-gp teaches us that transporter inhibition can involve far more complex mechanisms.

Conclusion

We were able to present three datasets concerning MRP2 inhibitors, inducers and substrates of reasonable size to apply computational methods on them. The models produced predicted their endpoint with high discriminatory power. Although the establishment of HTS screening for MRP2 could improve standardization and quantity of data available, we feel that decision trees can substantially contribute to extrude information even out of inhomogeneous data. An advantage of DTI is their rule-based output, which can easily be implemented without any specialized software. In this way, our models will, besides their contribution to our knowledge on MRP2 interaction, also serve as sensitive screening tools in drug development and toxicity screening.

Conflict of interest/disclosure

Claudia Suenderhauf is supported by the Swiss National Foundation (Grant No. 323530-119218).

5.5 A Binary Ant Colony Optimization Classifier for Molecular Activities

Felix Hammann [*], Claudia Suenderhauf, Jörg Huwyler

Division of Pharmaceutical Technology, Department of Pharmaceutical Sciences,
University of Basel, Basel, Switzerland

Corresponding Author (*):
Felix Hammann, MD, PhD
Department of Pharmaceutical Sciences
University of Basel
Klingelbergstrasse 50
4056 Basel, Switzerland
E-Mail: felix.hammann@unibas.ch
Phone: +41 61 267 15 13

Journal of Chemical Information and Modeling
2011 Sept 14.
(J. Chem. Inf. Model.)

Abstract

Chemical fingerprints encode the presence or absence of molecular features and are available in many large databases. Using a variation of the Ant Colony Optimization (ACO) paradigm, we describe a binary classifier based on feature selection from fingerprints. We discuss the algorithm and possible cross-validation procedures. As a real-world example, we use our algorithm to analyze a *Plasmodium falciparum* inhibition assay and contrast its performance with other machine learning paradigms in use today (decision tree induction, random forests, support vector machines, artificial neural networks). Our algorithm matches established paradigms in predictive power, yet supplies the medicinal chemist and basic researcher with easily interpretable results. Furthermore, models generated with our paradigm are easy to implement and can complement virtual screenings by additionally exploiting the pre-calculated fingerprint information.

Introduction

Chemical fingerprints, which are in essence hashes calculated from molecular structures, are frequently used in large chemical database.[309] These fixed-length strings encode a variety of molecular properties, oftentimes the presence or absence of a substructural motif.[310, 311] Using fingerprints, it is possible to quantify chemical similarity, restrict searches to a number of promising candidates, and so on.[312, 313] A variety of distance measures exist to determine proximity between two molecules and also to define clusters of similar structures.[309, 314]

According to the similarity principle, molecules with closely related structures are likely to exhibit the same activity.[315] While similarity may be defined as small distance from a selected point (e.g., a centroid in clustering), one may also construct a plane of separation within the attribute space to distinguish one type of molecule from another. A typical example of this approach are support vector machines,[316] and these have been successfully employed in many quantitative structure-activity relationship (QSAR) studies.[317]

Statistical models often rely on a number of physicochemical descriptors, which are rarely available in chemical databases. Screening an entire database therefore requires retrieving the complete set of structures and subsequently calculating these properties. On the other hand, fingerprints already contain many properties calculated upon insertion into the database. A classification scheme based on fingerprints could therefore save data traffic and computing power.

Such a classifier would need to find a subset of attributes present in a list of desirable molecules and absent in a list of negative controls. This feature selection is essentially an optimization problem. In recent years, the field of natural computing has produced intriguing heuristics for optimizations in engineering and the natural sciences. Ant Colony Optimization (ACO), a paradigm introduced in the 1990s,[224] has drawn a special amount of attention. Real-world ants are abstracted as agents able to traverse a graph while they deposit a pheromone whose intensity decays over time. An ant scurrying about the graph at random until it finds a food source initiates the process. It then returns to the starting point in a more or less direct trajectory. Other ants explore the graph and weigh their choices of route by previously deposited pheromones. Eventually, shorter (i.e., more efficient) paths will extrude a more intense signal and become points of convergence.

Here, we propose a binary classifier that uses an ACO variant to select relevant molecular fragments. Modifications of ACO have been proposed previously for variable selection and reduction of dimensionality.[318] They have also found application in the field of drug discovery, where they, however, were used in ensemble prediction settings (as feature reduction prior to e.g., linear regression [319, 320] or support vector machines [320] QSAR/QSPR studies of anti-HIV activity and human serum albumin binding activity, respectively). ACO is also often employed (along with other optimization paradigms) in protein ligand-docking studies.[321] While ACO has been applied to molecular binary classification (e.g., as an estimator of splitting criteria in decision tree induction),[322] we are not aware of its solitary use in fragment analysis.

The variant described by us can be visualized with the ant colony at the center and various single fingerprint flags as the vertices of edges of equal length radiating from the center, i.e. a complete bipartite graph $S_{1,n}$ where n is the number of fingerprint flags (Figure 36). From this, the method compiles a subset of flags that are associated with a given label or activity.

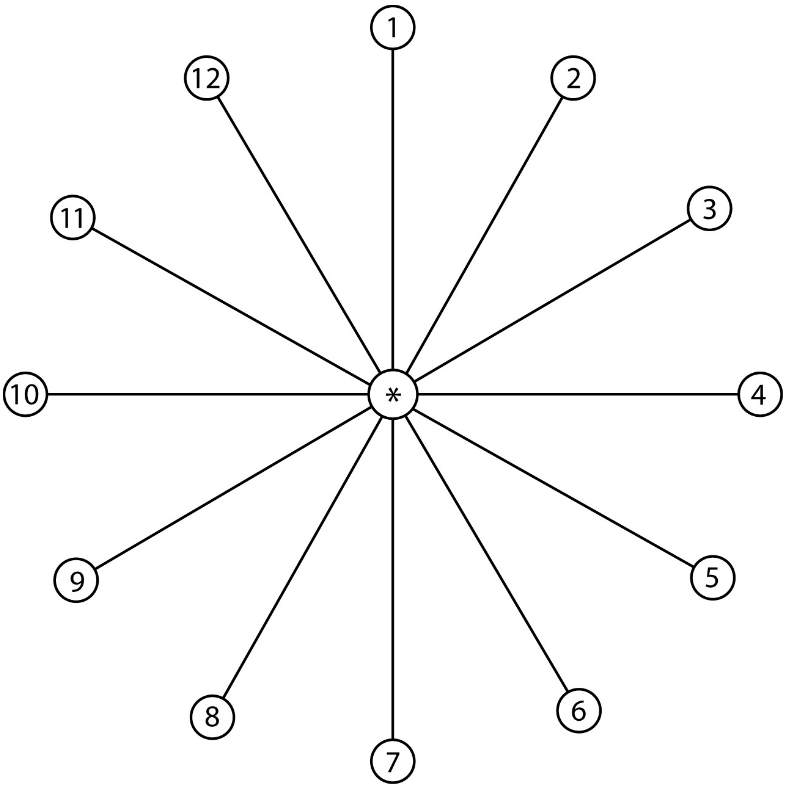


Figure 36 - Sample graphical depiction of the ant colony feature selection problem for twelve different attributes. Edges are of equal length, with the colony at the center (*).

Theory

Fingerprints

For a molecule *A*, a fingerprint *F* with *n* elements takes the form of an attribute vector F_A [309]

$$F_A = \{f_{1A}, f_{2A}, \ldots, f_{nA}\}$$

The fingerprints used in this context are dichotomous (or binary), i.e. each element f_{xA} corresponds to a bit with encodes the presence (ON) or absence (OFF) of a feature within the molecule *A*.

Classification problem

The problem can be stated as follows: given a total of n positions in a set F of fingerprints bits, find the subset S of magnitude m ($m<n, S \subset F$) so that the subset of features Q ($Q \subset F$) present in all active compounds has maximum specificity and sensitivity compared to the set of inactive compounds. Initially, a fixed number n_a of ants, each with the ability to select m features, explore the feature space at random, and return to the nest. This random exploration is implemented by assigning a random amount of pheromone τ (range 0 to 1) to each feature at the start of a run. Here, a heuristic fitness function H rates each ant's performance, and ants are ranked by quality of their subset. The best k ($k < n_a$) ants are selected and deposit a constant amount τ of pheromone on the edges connecting members of their subset to the nest. The other ants are ignored. All n edges are allowed to evaporate their pheromone trails by a constant linear term d, and a new cycle is initiated. Again, n_a ants are created. Each ant now generates a random number r_i for each i of the n edges connected to the nest ($0 \leq r_n \leq 1$) and ranks their attractiveness by choosing those with a maximal value for the term

$$a_i = r_i \tau_i$$

where τ_i is the intensity of the pheromone signal on edge i, and r_i the random weight. With the introduction of the random term, exploration of other combinations is encouraged. Over a given number of cycles, information-rich features are reinforced and increasingly become part of subsets until ants will almost uniformly choose these same features (Figure 37). Chart 1 further illustrates the proposed method.

```
KeyLength: length of fingerprint key in bits
m: number of features in ant memory
n: number of ants to create per cycle
k: number of ants to keep per cycle, where k <= n
Pheromones: list of length KeyLength containing pheromone intensities associated with keys

Initialize Pheromones with random floating point values from the range [0, 1]

repeat for a given number of cycles
    repeat for each of n ants
        copy Pheromones to Pheromones'
        multiply every position in Pheromones' with random floating point number from the
            range [0, 1]
        find m positions in Pheromones' with highest value
        compute fitness H and store with ant
    end repeat

    select k ants with highest H and repeat for each of m features in ant memory add constant
        pheromone amount $\tau$ to corresponding value in Pheromones

    for each position in Pheromones subtract constant linear evaporation rate
end repeat

Output: the list of pheromone intensities Pheromones
```

Chart 1 - Pseudocode representation of learning algorithm.

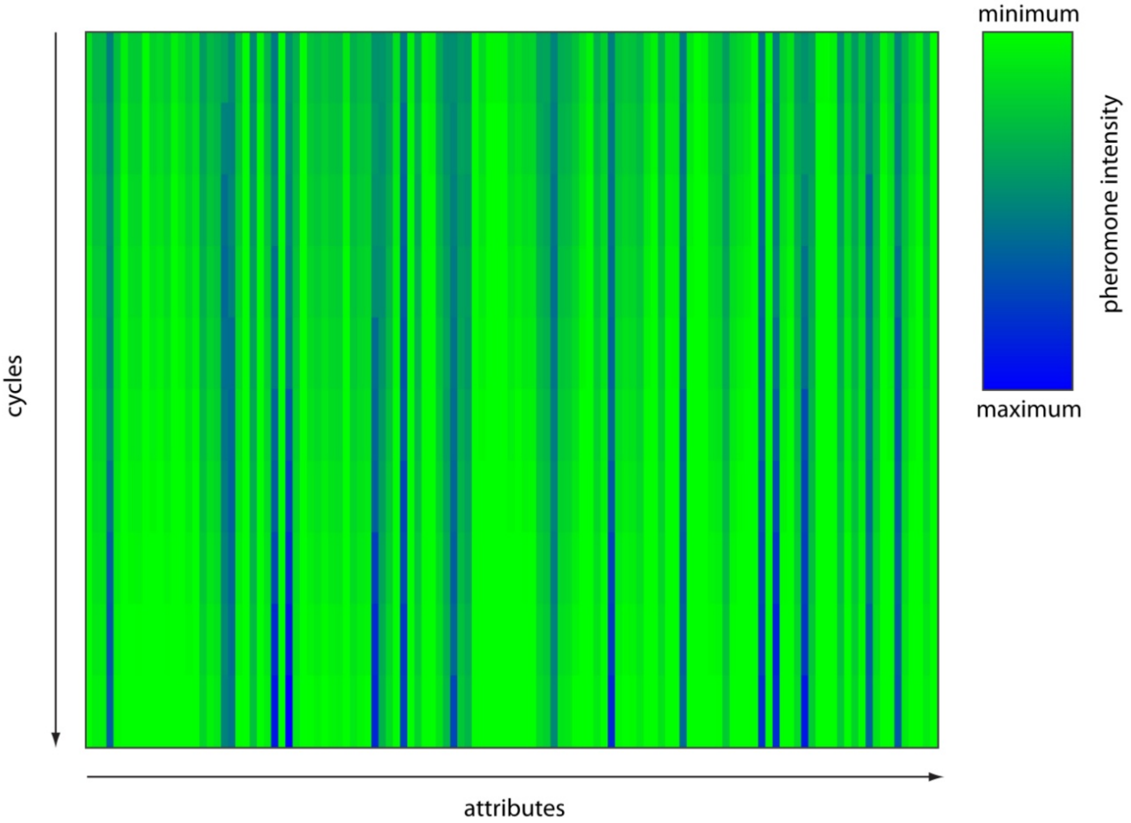


Figure 37 - Visualization of evolution of a solution over a number of cycles as produced by the software written for this study. Initially, many different attributes are being explored until several strong attributes are converged upon (indicated by intensity of pheromone trail).

Heuristic fitness function *H*

The heuristic fitness function *H* used in this study first finds the cardinality c_i of the intersection I_i of the subset *S* of features being evaluated and the entire set of features M_i ($M_i \subset F$) of each molecule *i* in a training dataset such that

$$I_i = S \cap M_i$$

Depending on the original parameterization of the ant agents, c_i will take on values between 0 and cardinality of *S*. The fitness function determines c_i for every instance in the training set and group instances by this value. Sensitivity and specificity of *S* for active molecules can be ranked by the area under the curve (AUC) of receiver operating characteristic (ROC) curves, as c_i as a cut-off value increases. This AUC is also the return value of *H(S)* for a subset *S*. A pseudocode representation is given in Chart 2.

```
TrainingData: training data with binary labels and fingerprint keys
AntMemory: set of m keys

repeat for each instance in TrainingData
    compute hits (i.e. $c_i$) as the number of keys present in both the instance and AntMemory
    sort and divide instances in TrainingData by hits
    continuously combining groups of instances ordered by value of hits, compute true and
        false positive rates for all instances as coordinates
    calculate the area under the curve formed by these points
end repeat
```

Chart 2 - Pseudocode representation of heuristic fitness function H.

Statement of models

The ROC curves are used further to determine the cut-off point with optimal sensitivity and specificity. The Youden index [323] J given as

$$J = \text{Sensitivity} + \text{Specificity} - 1$$

is maximal for this point (Section 4.5.3). A model built in this fashion can therefore be stated as the set of features indicative of activity and the minimum number of features required to qualify as active. A model *P* built with *m* features and cut-off point at *p* features takes the form of

$$P = \{\{x_1, x_2, \ldots, x_m\}, p\}$$

As an illustration, consider a sample model *M* trained from a set of 100 possible binary keys to select the 10 keys associated with a given activity. This might look as follows:

$$M = \{ \{ 4, 12, 15, 23, 38, 42, 61, 89, 90, 95 \}, 3 \}$$

For an instance to be classified as active, 3 or more of the 10 features would need to be present in its key vector, e.g., a molecule with a vector

$$\mathrm{Mol}_1 = \{ 10, \mathbf{12}, 19, 20, \mathbf{23}, \mathbf{38}, 49, 50, 67, 70, 82, 83, 100 \}$$

would classify as active (as the intersection with the key vector in model *M* has a cardinality of 3).

Cross-validation procedure

To avoid overfitting (i.e., creating overly complex models with very high predictive accuracy on training data by extracting too many parameters from the known data at the expense of not being able to predict unseen compounds), we used k-fold cross-validation (CV). Here, a data set is randomly recombined into k subsets (here k=10).[324] Of these, k-1 are re-combined to make up a training set which is tested against the remaining 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. Sets were permutated using the Fisher-Yates-Shuffle algorithm as detailed by Knuth.[325]

We evaluated three different ways of combining the different models: averaging of pheromone weights in every fold (averaged model), selection of most frequently employed attributes (frequency model), and combination of attributes most frequently selected by elite ants (elite ants model), i.e. the single best performing ant within a run. For averaging, the pheromone weights associated with each of the *n* attributes are normalized to a range (0, 1). Next, all of the *k* pheromones for a given attribute are summed up. The result is a list of *n* combined pheromone weights ranging from 0 to *k*, allowing them to be ranked. Attributes of the highest rank are selected and make up the final model. For the frequency model, the highest-ranking attribute of each fold is selected. In order to create the elite ants model, the software stores the single best performing ant of each fold. The corresponding feature sets are combined in the same manner as in the frequency model.

Performance measures

See Section 4.5.

Plasmodium falciparum growth inhibitor assay

Dataset and preparation

Models were learned from data of a high-throughput SYBR Green proliferation assay of P. falciparum (Pf) infected red blood cells published by Plouffe et al.[326] The data was retrieved from PubChem (http://pubchem.ncbi.nlm.nih.gov/) and contains a total of 1,272 compounds (201 active, 349 inactive, and 722 inconclusive). We omitted compounds labeled as inconclusive, as well as those for which not every CDK descriptor could be calculated (n=3). We removed disconnected small fragments such as counterions prior to any calculation in analogy to McGregor and Pallai.[310]

We evaluated three fingerprint keys (MACCS (MDL), STANDARD, EXTENDED) available in the latest stable Chemical Development Kit (Version 1.2.7).[327] The 166 bit MDL key [310] was used in the final model as it is the best documented of the three and has been optimized to allow for clustering of bioactive substances in the context of drug discovery. The concept of fingerprint darkness refers to the fraction of bits set to ON, i.e. we consider fingerprints with more bits set to ON as darker. The characteristics of the three different keys evaluated are given in Table 27.

Key	Class	mean%	max%	min%	sd%
MACCS	negative	29.7	49.4	4.8	7.8
	positive	24.1	49.4	1.2	9.2
	total	26.1	49.4	1.2	9.1
Standard	negative	9.9	49.9	0.1	8.3
	positive	17.1	60.1	0.8	9.6
	total	12.5	60.1	0.1	9.4
Extended	negative	10.2	51.4	0.1	8.4
	positive	17.6	59.7	0.8	9.6
	total	12.9	59.7	0.1	9.6

Table 27 - Mean fingerprint darkness (number of bits set over total number of bits), with minimum (min%), maximum (max%) percentages and standard deviation in percent (sd%) of the 166 bit MACCS key and the 1024 bit standard and extended keysets.

Of these, the MACCS 166 bit key shows the greatest darkness (26.1%), implying that it is capable of reflecting the most features with the least computational effort.

Training of models

We let the classifier learn over 100 cycles with to produce models of with a magnitude of 10. Ants deposited a pheromone amount τ = 0.1 which evaporated by d = 0.05 within cycles. We performed 100 runs using the three different modes of cross-validation outlined above. This amounted to a total of 300 models. For comparison, we created models with a decision tree induction algorithm (J4.8, a C4.5 variant), random forests (RF) of ten trees with five attributes each,[328] support vector machines (SVM) using a polynomial kernel function, and artificial neural networks (ANN) with a single hidden layer. In line with other current studies, models were learned in a 10-fold cross-validated context.[260, 329] The numerical attributes used in this process were the 1D and 2D descriptors (n = 27) available in the CDK (molecular weight, calculated partitioning coefficient (LogP), topological polar surface area, BCUT metrics, fragment complexity, atom and bond counts of aromatic and of all atoms, hydrogen bond donor and acceptor counts, Kier-Hall shape indices, Petitjean number, number of rotatable bonds, atomic polarizability, length of largest chain and largest aliphatic chain, as well as length of largest π chain).[330]

Model performance

The classification results of the best performing models for each mode of cross-validation are shown in Table 28. All three CV procedures achieve comparable CCRs of 0.84 to 0.87 - values that match those of the other paradigms implemented. An elite ant model achieved the highest CCR (0.87). Its associated ROC curve has a high area under the curve of 0.91 and is given in Figure 38. It is readily apparent from Table 29, which presents the substructural motifs selected by the model, that the binary ACO classifier retrieves fragments with a mechanistical relevance.

	TP	TN	FN	FP	CCR	MCC	Accuracy
Elite ACO Model (Run 36)	171	307	29	40	0.87	0.73	0.87
Frequency ACO Model (Run 17)	169	294	31	53	0.85	0.68	0.85
Averaged ACO Model (Run 16)	162	303	38	44	0.84	0.68	0.85
J48	156	313	44	34	0.84	0.69	0.86
RF	155	316	45	31	0.84	0.70	0.86
SVM	157	317	43	30	0.85	0.71	0.87
ANN	161	296	39	51	0.83	0.65	0.84

Table 28 - Results of binary ant colony optimization (ACO) classification using three different cross-validation paradigms and comparison with established machine learning paradigms. The data consists of 547 instances (positive: 200, negative: 347). J48: decision tree induction, RF: random forests, SVM: support vector machines, ANN: artificial neural networks. Performance is measured as corrected classification rate (CCR), Matthews correlation coefficient (MCC), and accuracy.

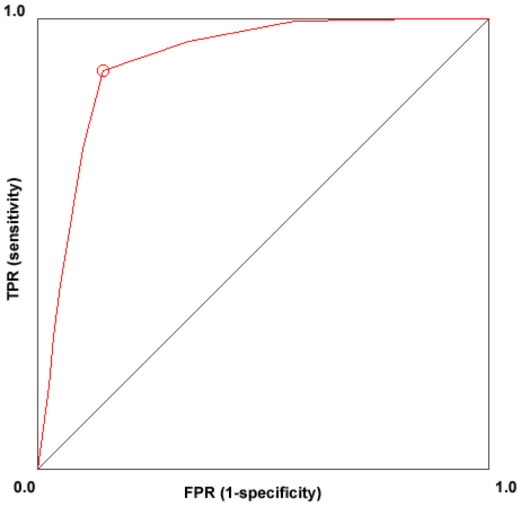


Figure 38 - Receiver operating characteristic curve for the best performing classification model in this study (area under the curve = 0.91). The circle denotes the cut-off point from which on instances are classified as positive.

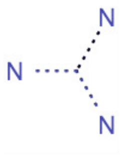
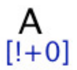
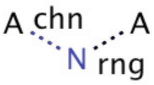
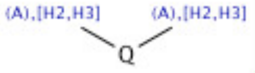
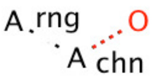
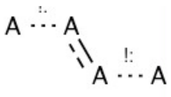
Index	Depiction	SMARTS	Comment
25		[#7]~[#6](~[#7])~[#7]	trisamino / imino methylene.
49		[!+0]	presence of charge.
75		*!@[#7]@*	interposition of nitrogen.
86		[C;H2,H3][!#6;!#1][C;H2,H3]	carbon – heteroatom – carbon chain.
124		[!#6;!#1]~[!#6;!#1]	two connected heteroatoms.
127		*@*!@[#8]	oxygen connected to any ring system via a single bond.
137		[!C;!c;R]	any heterocycle.
140	O	[#8]	presence of oxygen.
144		*!:*:*!:*	aromatic ring substituted in ortho-position by two non-aromatic substituents.
161	N	[#7]	presence of nitrogen.

Table 29 - Fingerprint keys associated with a Plasmodium falciparum growth inhibition as determined by binary ant colony optimization classification. Substructural motifs are given with their position (index), SMILES arbitrary target specification (SMARTS) along with an image and an explanation.

For instance, the presence of nitrogen and oxygen atoms in different frameworks (keys 25, 75, 127, 140, and 161) is characteristic of drug-like molecules (hydrogen bonding capacity) as well as important for successfully overcoming cellular membranes. They are also present in molecules which exert oxidative stress, to which *P. falciparum* is very sensitive. The prevalence of both nitrogen and oxygen are high in small molecule drugs. Of course, some rare examples exist which contain neither (noteworthy members of this class are lindane and mitotane, two chemotherapeutic agents) and many molecules contain exclusively one these atom species (e.g., nitrogen in amitryptiline, selegeline, and memantine, and oxygen in ivermectin, digoxin, and cholecalciferol). This shows how the substructures identified by the paradigm need to be interpreted together. For example, nitrogen appears in other keys (keys 25 (trisamino / imino methylene) and 75). Presence of one of these more complex substructures therefore automatically increases the score and, by consequence, the likelihood of positive classification. In the same vein, molecules lacking keys 25 or 75 can improve their score by offering other hydrogen binding sites, e.g., oxygen (keys 140 and 127), thereby increasing their drug-likeness.

A key enzyme in the life cycle of Pf, the cysteine protease Falcipain-2 (FP-2) that degrades hemoglobin (Hb), can be inhibited by certain epoxysuccinates and aziridinyl substituents in quinone rings (perceived, amongst others, by keys 86 and 137) have been shown to enhance antiplasmodial activity by inhibiting Pf glutathion reductase.[331] The life cycle of Pf is particularly vulnerable during the erythrocytic stage as its metabolism is largely anaerobic and hence sensitive to oxidative stress.

Conclusions

We investigated whether binary classification of molecular activity using a variation of the ACO paradigm could become a valid alternative to other ML classification methodologies. Analysis of the Pf inhibition assay by Plouffe et al. shows the high degrees of accuracy achieved by our models and their competitiveness with established ML methods.[326]

The different modes of CV produce similarly powerful classifiers. From Table 28 it is evident that these models stem from different runs, i.e. the choice of CV influences the final performance, and no final ranking can be made between these modes. Therefore, we consider it advisable to calculate all three to maximally exploit the information extracted by the learning process.

The information provided to the binary ACO learning algorithm was in essence a list of the presence or absence of substructural motifs or fragments, i.e. two-dimensional structural information. We, therefore, explicitly learned the alternative ML methods from two-dimensional descriptors as well to ensure a level playing field. Arguably, one might see better performance of the established ML methods with a different choice of descriptors. Conversely, other fingerprint keys could improve the results of binary ACO classification.

We chose MACCS over the other available fingerprint keys in CDK because of its length (166 bits vs. 1024 bits for Standard and Extended) and its high ratio of keys set to on. Notably, the molecules tagged as negative have a higher fingerprint darkness than the positive instances, i.e., the inactive compounds are actually captured better than the active ones. When one learns a model to distinguish active compounds by presence of certain features from such a data set, it is apparent that the algorithm cannot simply associate fingerprint darkness of a compound with activity.

The substructures encoded in the MACCS fingerprint are oftentimes ambiguous or very general, and features selected by our algorithm can overlap (e.g., keys 25 and 161). Still, models perform well and robustly in a cross-validated setting. This indicates that the subsets of keys are more than the sum of their parts, i.e. the individual contribution of a key must be seen in the context of the entire subset. Also, a feature that is recognized by several keys is amplified (or deemed more important) in the perception of the classifier.

Binary ACO models can benefit the drug discovery process in two principle ways. First, the models provide an explicit fragment analysis directly accessible to human interpretation. Medicinal chemists can use them as guides for further development. Secondly, models can be applied directly to existing databases without any further calculations if both use the same fingerprinting scheme. This is in contrast to more elaborate numerical methods (e.g., SVM or ANN) where a) a number of physicochemical descriptors need to be computed, and b) the software implementation of the classifier itself is complex. In fact, binary ACO models learned from fingerprints could be implemented as native database queries.

Of the learning paradigms employed in this study, decision tree induction took the least time to produce models. This, of course, does not consider the time required for calculating descriptors, performing intercorrelation analysis, and checking for missing values. Support vector machines had the most time-intensive learning process. For SVM, we are not considering the tedious process of optimizing learning parameters. Similar considerations, of course, have also to be made when applying the binary ACO algorithm. Proper choice of the model size, *m*, influences not only the interpretability and performance on unseen data, but also the time required to learn models. The number of cycles being spent on learning contributes directly to the computational expense. In summing up, the proposed algorithm ranks with SVMs in terms of time consumption for learning. A more thorough profiling does not seem called for, as the learning paradigms differ in practice in the amount of data preparation and optimization they require.

In the future, this algorithm could be extended to numerical predictions, i.e. the learning process could correlate number of keys present with the degree of activity. Additionally, instead of merely identifying keys contributing to activity, a variant of the algorithm proposed here might single out detrimental features and incorporate them in the predictions.

Software used

See Section 4.6.3.
Feature selection was performed using in-house software.

Acknowledgments

The authors wish to thank Dr Heinz Hammann for the fruitful discussions regarding CDK fingerprints. Dr. Claudia Suenderhauf is supported by the Swiss National Foundation (Grant No. 323530-119218).

Supporting Information

An extended table gives more structural information on the molecules analyzed in this study. This material is available free of charge via the Internet at http://pubs.acs.org. (See Section 7.4.)

6 Conclusion and outlook

It was the aim of the present thesis to assess drug transport across several physiological barriers using computational methods. As these barriers are found on different sites in the body such as the intestinal wall, the CNS, and the lactating breast epithelium, they vary in their characteristics and degree of tightness. However, all of them are able to alter drug distribution and pharmacokinetics. Generally, one could use *in vitro* methods to gain information on pharmacokinetics on the cellular level. However, these models are limited as they do not reflect complexity of the living system. Therefore, pharmacokinetic studies are ideally performed *in vivo*. Both methods can be time consuming and expensive, which does not meet the requirements for fast screening of chemical libraries. Meeting the demand for efficient ways to assess thousands of compounds for their pharmacokinetic behavior, *in silico* methods were developed to predict several relevant endpoints. The information content, which these models hold in turn shed new light on the processes involved. In the present thesis, it could indeed be shown that pharmacokinetic modeling can be applied to different barriers:

In the first study, robust and accurate models were presented to predict human intestinal absorption. Emphasis was put on comparing different feature sets and performance of various machine learning paradigms. Although a variety of approaches was used, models revealed uniformly well-known features, such as measures of charge and lipophilicity, but also descriptors which are less commonly used to model human intestinal absorption, such as structural symmetry.

Models for drug brain penetration reconfirmed the importance of well-known physicochemical features from other models, such as lipophilicity, size, and charge. However, substructure analysis and decision trees added new perspectives for predicting brain penetration, such as the involvement of stereochemistry. The underlying data were based on experimental LogPS values retrieved from rats. Although LogPS data of mice were available, they were not included into the present models to avoid bias due to relevant effects in different species. Ideally, a prediction for drug brain penetration would be based on data retrieved in humans. Hopefully, such data could be acquired in the future with non-invasive techniques.

Data preparation for the models dealing with prediction of drug permeation from maternal plasma into breast milk was a challenging endeavor. The data underlying these models was retrieved from literature, where information was available from nursing mothers. However, the quantities of drug retrieved in breast milk were often reported in an ambiguous manner. In order to produce meaningful and predictive models, numerous drugs with ambiguous endpoints had to be excluded from learning. Although it would be desirable to have a bigger data source of numeric data, models created on these imprecise endpoints performed with exceptional accuracy. Breast milk varies in nutrient composition regarding proteins and fat content during the lactating period. It is most likely that this change in composition would also have impacts on drug permeation. Future studies might address this issue by establishing models for single lactating periods.

In the study of MRP2 substrates, inducers and inhibitors, three new datasets retrieved from literature. Decision tree models of inducer and inhibitors data, as well as the substructure search for substrates, revealed insights in the requirements for MRP2 interaction. However, for these endpoints no gold standard for experimental settings and data acquisition has yet been defined. The complexity of influx, metabolic enzyme, and MRP2 interplay hampered the development of robust *in vitro* models, which would qualify for high-throughput screening. While the cross-validated models presented here performed strongly, models based on uniformly acquired data would probably yield even more information.

In our last project, an ant colony optimization algorithm was presented to perform fragment based feature selection. The paradigm was tested on the highly combined endpoint of predicting drugs with antimalarial activity. The chemical fingerprints selected gave direct information on structural requirements for drug activity, without distinguishing different action mechanisms. By implementing an extension for numerical predictions, fingerprints could be correlated with different levels of activity. The paradigm could be additionally enhanced to make statements on structural requirements for inactive compounds.

The accuracy and performance the models presented here is encouraging and shows that pharmacokinetic endpoints can be successfully assessed by computational methods. The adaption of new machine learning techniques and advances in data acquisition will therefore offer additional perspectives for *in silico* methods.

7 Appendix

7.1 Bibliography

1. Solecki, R.S., Shanidar, the first flower people. 1st ed. 1971, New York: Knopf. xv, 290, x p.

2. Schlegel, B., Luhmann, U., Hartl, A., and Grafe, U., Piptamine, a new antibiotic produced by Piptoporus betulinus Lu 9-1. J. Antibiot. (Tokyo), 2000. **53**(9): p. 973-4.

3. Kamo, T., Asanoma, M., Shibata, H., and Hirota, M., Anti-inflammatory lanostane-type triterpene acids from Piptoporus betulinus. J. Nat. Prod., 2003. **66**(8): p. 1104-6.

4. Fowler, B., Iceman : uncovering the life and times of a prehistoric man found in an alpine glacier. University of Chicago Press ed. 2001, Chicago: University of Chicago Press. xv, 315 p.

5. Koshland, D.E., Application of a Theory of Enzyme Specificity to Protein Synthesis. Proc. Natl. Acad. Sci. USA, 1958. **44**(2): p. 98-104.

6. Campbell, W.C., Fisher, M.H., Stapley, E.O., Albers-Schonberg, G., and Jacob, T.A., Ivermectin: a potent new antiparasitic agent. Science, 1983. **221**(4613): p. 823-8.

7. Borel, J.F., History of the discovery of cyclosporin and of its early pharmacological development. Wien. Klin. Wochenschr., 2002. **114**(12): p. 433-7.

8. Drews, J., Drug discovery: a historical perspective. Science, 2000. **287**(5460): p. 1960-4.

9. Lipinski, C.A., Lombardo, F., Dominy, B.W., and Feeney, P.J., Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv. Drug Deliv. Rev., 2001. **46**(1-3): p. 3-26.

10. Macheras, P., Reppas, C., and Dressman, J.B., Biopharmaceutics of orally administred dosage forms. Vol. 1. 1995, Chichester, UK: Ellis Horwood Ltd.

11. Wermeling, D.P., Intranasal delivery of antiepileptic medications for treatment of seizures. Neurotherapeutics, 2009. **6**(2): p. 352-8.

12. Evans, H.C. and Easthope, S.E., Transdermal buprenorphine. Drugs, 2003. **63**(19): p. 1999-2010; discussion 2011-2.

13. Godoy, C., Greenspahn, B.R., Kushner, M.J., Lahti, R.E., Levy, R.A., Reese, P., Reynolds, G.M., Jr., and Rice, S.C., Comparison of two transdermal nitroglycerin systems. Clin. Ther., 1986. **8**(6): p. 689-93.

14. Nicholson, B.P. and Schachat, A.P., A review of clinical trials of anti-VEGF agents for diabetic retinopathy. Graefes Arch. Clin. Ex.p Ophthalmol., 2010. **248**(7): p. 915-30.

15. Norinder, U., Osterberg, T., and Artursson, P., Theoretical calculation and prediction of Caco-2 cell permeability using MolSurf parametrization and PLS statistics. Pharm. Res., 1997. **14**(12): p. 1786-91.

16. Winiwarter, S., Ax, F., Lennernas, H., Hallberg, A., Pettersson, C., and Karlen, A., Hydrogen bonding descriptors in the prediction of human in vivo intestinal permeability. J. Mol. Graph. Model., 2003. **21**(4): p. 273-87.

17. Kerns, E.H. and Di, L., Drug-like properties: concepts, structure design and methods. From ADME to toxicity optimization. Vol. 1. 2008: Academic Press.

18. Lipinski, C.A., Drug-like properties and the causes of poor solubility and poor permeability. J. Pharmacol. Toxicol. Methods, 2000. **44**(1): p. 235-49.

19. Artursson, P., Epithelial transport of drugs in cell culture. I: A model for studying the passive diffusion of drugs over intestinal absorptive (Caco-2) cells. J. Pharm. Sci., 1990. **79**(6): p. 476-82.

20. Artursson, P. and Magnusson, C., Epithelial transport of drugs in cell culture. II: Effect of extracellular calcium concentration on the paracellular transport of drugs of different lipophilicities across monolayers of intestinal epithelial (Caco-2) cells. J. Pharm. Sci., 1990. **79**(7): p. 595-600.

21. Fagerholm, U., Nilsson, D., Knutson, L., and Lennernas, H., Jejunal permeability in humans in vivo and rats in situ: investigation of molecular size selectivity and solvent drag. Acta Physiol. Scand., 1999. **165**(3): p. 315-24.

22. Oswald, S., Terhaag, B., and Siegmund, W., In vivo probes of drug transport: commonly used probe drugs to assess function of intestinal P-glycoprotein (ABCB1) in humans. Handb. Exp. Pharmacol., 2011(201): p. 403-47.

23. Huwyler, J., Wright, M.B., Gutmann, H., and Drewe, J., Induction of cytochrome P450 3A4 and P-glycoprotein by the isoxazolyl-penicillin antibiotic flucloxacillin. Curr. Drug Metab., 2006. **7**(2): p. 119-26.

24. Fricker, G., Drewe, J., Huwyler, J., Gutmann, H., and Beglinger, C., Relevance of p-glycoprotein for the enteral absorption of cyclosporin A: in vitro-in vivo correlation. Br. J. Pharmacol., 1996. **118**(7): p. 1841-7.

25. Dahan, A. and Amidon, G.L., Segmental dependent transport of low permeability compounds along the small intestine due to P-glycoprotein: the role of efflux transport in the oral absorption of BCS class III drugs. Mol. Pharm., 2009. **6**(1): p. 19-28.

26. Benet, L.Z., The drug transporter-metabolism alliance: uncovering and defining the interplay. Mol. Pharm., 2009. **6**(6): p. 1631-43.

27. Christians, U., Schmitz, V., and Haschke, M., Functional interactions between P-glycoprotein and CYP3A in drug metabolism. Expert Opin. Drug Metab. Toxicol., 2005. **1**(4): p. 641-54.

28. Mak, M., Fung, L., Strasser, J.F., and Saltzman, W.M., Distribution of drugs following controlled delivery to the brain interstitium. J. Neurooncol., 1995. **26**(2): p. 91-102.

29. Jain, R.K., Tumor physiology and antibody delivery. Front. Radiat. Ther. Oncol., 1990. **24**: p. 32-46; discussion 64-8.

30. Liu, X. Factors affecting total and free drug concentration in the brain. in AAPS Conference: critical issues in discovering quality clinical candidates. 2006. Philadelphia, PA.

31. Abbott, N.J., Ronnback, L., and Hansson, E., Astrocyte-endothelial interactions at the blood-brain barrier. Nat. Rev. Neurosci., 2006. **7**(1): p. 41-53.

32. Pardridge, W.M., CNS drug design based on principles of blood-brain barrier transport. J. Neurochem., 1998. **70**(5): p. 1781-92.

33. Clark, D.E., In silico prediction of blood-brain barrier permeation. Drug Discov. Today, 2003. **8**(20): p. 927-33.

34. Platts, J.A., Abraham, M.H., Zhao, Y.H., Hersey, A., Ijaz, L., and Butina, D., Correlation and prediction of a large blood-brain distribution data set-an LFER study. Eur. J. Med. Chem., 2001. **36**(9): p. 719-30.

35. Lanevskij, K., Japertas, P., Didziapetris, R., and Petrauskas, A., Ionization-specific QSAR models of blood-brain penetration of drugs. Chem. Biodivers., 2009. **6**(11): p. 2050-4.

36. Bendels, S., Kansy, M., Wagner, B., and Huwyler, J., In silico prediction of brain and CSF permeation of small molecules using PLS regression models. Eur. J. Med. Chem., 2008. **43**(8): p. 1581-92.

37. Goodwin, J.T. and Clark, D.E., In silico predictions of blood-brain barrier penetration: considerations to "keep in mind". J. Pharmacol. Exp. Ther., 2005. **315**(2): p. 477-83.

38. Van Asperen, J., Schinkel, A.H., Beijnen, J.H., Nooijen, W.J., Borst, P., and van Tellingen, O., Altered pharmacokinetics of vinblastine in Mdr1a P-glycoprotein-deficient Mice. J. Natl. Cancer. Inst., 1996. **88**(14): p. 994-9.

39. Schinkel, A.H., Wagenaar, E., van Deemter, L., Mol, C.A., and Borst, P., Absence of the mdr1a P-Glycoprotein in mice affects tissue distribution and pharmacokinetics of dexamethasone, digoxin, and cyclosporin A. J. Clin. Invest., 1995. **96**(4): p. 1698-705.

40. Cordon-Cardo, C., O'Brien, J.P., Casals, D., Rittman-Grauer, L., Biedler, J.L., Melamed, M.R., and Bertino, J.R., Multidrug-resistance gene (P-glycoprotein) is expressed by endothelial cells at blood-brain barrier sites. Proc. Natl. Acad. Sci. USA, 1989. **86**(2): p. 695-8.

41. Schinkel, A.H. and Jonker, J.W., Mammalian drug efflux transporters of the ATP binding cassette (ABC) family: an overview. Adv. Drug Deliv. Rev., 2003. **55**(1): p. 3-29.

42. Poller, B., Drewe, J., Krahenbuhl, S., Huwyler, J., and Gutmann, H., Regulation of BCRP (ABCG2) and P-glycoprotein (ABCB1) by cytokines in a model of the human blood-brain barrier. Cell Mol Neurobiol, 2010. **30**(1): p. 63-70.

43. Cooray, H.C., Blackmore, C.G., Maskell, L., and Barrand, M.A., Localisation of breast cancer resistance protein in microvessel endothelium of human brain. Neuroreport, 2002. **13**(16): p. 2059-63.

44. Dauchy, S., Dutheil, F., Weaver, R.J., Chassoux, F., Daumas-Duport, C., Couraud, P.O., Scherrmann, J.M., De Waziers, I., and Decleves, X., ABC transporters, cytochromes P450 and their main transcription factors: expression at the human blood-brain barrier. J. Neurochem., 2008. **107**(6): p. 1518-28.

45. Lin, J.H. and Yamazaki, M., Clinical relevance of P-glycoprotein in drug therapy. Drug Metab. Rev., 2003. **35**(4): p. 417-54.

46. Agarwal, S., Sane, R., Gallardo, J.L., Ohlfest, J.R., and Elmquist, W.F., Distribution of gefitinib to the brain is limited by P-glycoprotein (ABCB1) and breast cancer resistance protein (ABCG2)-mediated active efflux. J. Pharmacol. Exp. Ther., 2010. **334**(1): p. 147-55.

47. De Vries, N.A., Zhao, J., Kroon, E., Buckle, T., Beijnen, J.H., and van Tellingen, O., P-glycoprotein and breast cancer resistance protein: two dominant transporters working together in limiting the brain penetration of topotecan. Clin. Cancer. Res., 2007. **13**(21): p. 6440-9.

48. Poller, B., Wagenaar, E., Tang, S.C., and Schinkel, A.H., Double-transduced MDCKII cells to study human P-glycoprotein (ABCB1) and breast cancer resistance protein (ABCG2) interplay in drug transport across the blood-brain barrier. Mol. Pharm., 2011. **8**(2): p. 571-82.

49. Varma, M.V., Ambler, C.M., Ullah, M., Rotter, C.J., Sun, H., Litchfield, J., Fenner, K.S., and El-Kattan, A.F., Targeting intestinal transporters for optimizing oral drug absorption. Curr. Drug Metab., 2010. **11**(9): p. 730-42.

50. Ogihara, T., Kano, T., Wagatsuma, T., Wada, S., Yabuuchi, H., Enomoto, S., Morimoto, K., Shirasaka, Y., Kobayashi, S., and Tamai, I., Oseltamivir (tamiflu) is a substrate of peptide transporter 1. Drug Metab. Dispos., 2009. **37**(8): p. 1676-81.

51. Pardridge, W.M., Log(BB), PS products and in silico models of drug brain penetration. Drug Discov. Today, 2004. **9**(9): p. 392-3.

52. Cunningham, A.S., Jelliffe, D.B., and Jelliffe, E.F., Breast-feeding and health in the 1980s: a global epidemiologic review. J. Pediatr., 1991. **118**(5): p. 659-66.

53. Walker, A., Breast milk as the gold standard for protective nutrients. J. Pediatr., 2010. **156**(2): p. 3-7.

54. Almroth, S., Greiner, T., and Latham, M.C., Economic importance of breastfeeding. Food Nutr. (Roma), 1979. **5**(2): p. 4-10.

55. Chandra, R.K., Prospective studies of the effect of breast feeding on incidence of infection and allergy. Acta Paediatr. Scand., 1979. **68**(5): p. 691-4.

56. Alberts, E., Kalverboer, A.F., and Hopkins, B., Mother-infant dialogue in the first days of life: an observational study during breast-feeding. J. Child Psychol. Psychiatry, 1983. **24**(1): p. 145-61.

57. Usher, K. and Foster, K., The use of psychotropic medications with breastfeeding women: applying the available evidence. Contemp. Nurse., 2006. **21**(1): p. 94-102.

58. Stultz, E.E., Stokes, J.L., Shaffer, M.L., Paul, I.M., and Berlin, C.M., Extent of medication use in breastfeeding women. Breastfeed. Med., 2007. **2**(3): p. 145-51.

59. Leung, B.M. and Kaplan, B.J., Perinatal depression: prevalence, risks, and the nutrition link-a review of the literature. J. Am. Diet. Assoc., 2009. **109**(9): p. 1566-75.

60. Spencer, J.P., Gonzalez, L.S., and Barnhart, D.J., Medications in the breast-feeding mother. Am. Fam. Physician, 2001. **64**(1): p. 119-26.

61. Begg, E.J., Atkinson, H.C., and Duffull, S.B., Prospective evaluation of a model for the prediction of milk:plasma drug concentrations from physicochemical characteristics. Br. J. Clin. Pharmacol., 1992. **33**(5): p. 501-5.

62. Begg, E.J. and Atkinson, H.C., Modelling of the passage of drugs into milk. Pharmacol. Ther., 1993. **59**(3): p. 301-10.

63. Atkinson, H.C. and Begg, E.J., Prediction of drug distribution into human milk from physicochemical characteristics. Clin. Pharmacokinet., 1990. **18**(2): p. 151-67.

64. McNamara, P.J., Meece, J.A., and Paxton, E., Active transport of cimetidine and ranitidine into the milk of Sprague Dawley rats. J. Pharmacol. Exp. Ther., 1996. **277**(3): p. 1615-21.

65. Schadewinkel-Scherkl, A.M., Rasmussen, F., Merck, C.C., Nielsen, P., and Frey, H.H., Active transport of benzylpenicillin across the blood-milk barrier. Pharmacol. Toxicol., 1993. **73**(1): p. 14-9.

66. Oo, C.Y., Kuhn, R.J., Desai, N., and McNamara, P.J., Active transport of cimetidine into human milk. Clin. Pharmacol. Ther., 1995. **58**(5): p. 548-55.

67. Mani, O., Korner, M., Ontsouka, C.E., Sorensen, M.T., Sejrsen, K., Bruckmaier, R.M., and Albrecht, C., Identification of ABCA1 and ABCG1 in milk fat globules and mammary cells-implications for milk cholesterol secretion. J. Dairy Sci., 2011. **94**(3): p. 1265-76.

68. Gilchrist, S.E. and Alcorn, J., Lactation stage-dependent expression of transporters in rat whole mammary gland and primary mammary epithelial organoids. Fundam Clin Pharmacol, 2010. **24**(2): p. 205-14.

69. Van Herwaarden, A.E. and Schinkel, A.H., The function of breast cancer resistance protein in epithelial barriers, stem cells and milk secretion of drugs and xenotoxins. Trends Pharmacol. Sci., 2006. **27**(1): p. 10-6.

70. Jonker, J.W., Merino, G., Musters, S., van Herwaarden, A.E., Bolscher, E., Wagenaar, E., Mesman, E., Dale, T.C., and Schinkel, A.H., The breast cancer resistance protein BCRP (ABCG2) concentrates drugs and carcinogenic xenotoxins into milk. Nat. Med., 2005. **11**(2): p. 127-9.

71. Van Herwaarden, A.E., Wagenaar, E., Merino, G., Jonker, J.W., Rosing, H., Beijnen, J.H., and Schinkel, A.H., Multidrug transporter ABCG2/breast cancer resistance protein secretes riboflavin (vitamin B2) into milk. Mol. Cell. Biol., 2007. **27**(4): p. 1247-53.

72. Ito, S. and Koren, G., A novel index for expressing exposure of the infant to drugs in breast milk. Br. J. Clin. Pharmacol., 1994. **38**(2): p. 99-102.

73. Leake, R.D. and Trygstad, C.W., Glomerular filtration rate during the period of adaptation to extrauterine life. Pediatr. Res., 1977. **11**(9 Pt 1): p. 959-62.

74. Rosati, A., Maniori, S., Decorti, G., Candussio, L., Giraldi, T., and Bartoli, F., Physiological regulation of P-glycoprotein, MRP1, MRP2 and cytochrome P450 3A2 during rat ontogeny. Dev. Growth Differ., 2003. **45**(4): p. 377-87.

75. Silverman, W.A., The schizophrenic career of a "monster drug". Pediatrics, 2002. **110**(2 Pt 1): p. 404-6.

76. Maliepaard, M., Scheffer, G.L., Faneyte, I.F., van Gastelen, M.A., Pijnenborg, A.C., Schinkel, A.H., van De Vijver, M.J., Scheper, R.J., and Schellens, J.H., Subcellular localization and distribution of the breast cancer resistance protein transporter in normal human tissues. Cancer Res, 2001. **61**(8): p. 3458-64.

77. He, G., Massarella, J., and Ward, P., Clinical pharmacokinetics of the prodrug oseltamivir and its active metabolite Ro 64-0802. Clin. Pharmacokinet., 1999. **37**(6): p. 471-84.

78. Netsch, M.I., Gutmann, H., Schmidlin, C.B., Aydogan, C., and Drewe, J., Induction of CYP1A by green tea extract in human intestinal cell lines. Planta Med., 2006. **72**(6): p. 514-20.

79. Gutmann, H., Poller, B., Buter, K.B., Pfrunder, A., Schaffner, W., and Drewe, J., Hypericum perforatum: which constituents may induce intestinal MDR1 and CYP3A4 mRNA expression? Planta Med., 2006. **72**(8): p. 685-90.

80. Miksys, S. and Tyndale, R.F., Brain drug-metabolizing cytochrome P450 enzymes are active in vivo, demonstrated by mechanism-based enzyme inhibition. Neuropsychopharmacology, 2009. **34**(3): p. 634-40.

81. Meyer, R.P., Gehlhaus, M., Knoth, R., and Volk, B., Expression and function of cytochrome p450 in brain drug metabolism. Curr. Drug Metab., 2007. **8**(4): p. 297-306.

82. Ghosh, C., Gonzalez-Martinez, J., Hossain, M., Cucullo, L., Fazio, V., Janigro, D., and Marchi, N., Pattern of P450 expression at the human blood-brain barrier: roles of epileptic condition and laminar flow. Epilepsia, 2010. **51**(8): p. 1408-17.

83. Dauchy, S., Miller, F., Couraud, P.O., Weaver, R.J., Weksler, B., Romero, I.A., Scherrmann, J.M., De Waziers, I., and Decleves, X., Expression and transcriptional regulation of ABC transporters and cytochromes P450 in hCMEC/D3 human cerebral microvascular endothelial cells. Biochem. Pharmacol., 2009. **77**(5): p. 897-909.

84. Baron, J., Voigt, J.M., Whitter, T.B., Kawabata, T.T., Knapp, S.A., Guengerich, F.P., and Jakoby, W.B., Identification of intratissue sites for xenobiotic activation and detoxication. Adv. Exp. Med. Biol., 1986. **197**: p. 119-44.

85. Krishna, D.R. and Klotz, U., Extrahepatic metabolism of drugs in humans. Clin. Pharmacokinet., 1994. **26**(2): p. 144-60.

86. Neafsey, P., Ginsberg, G., Hattis, D., and Sonawane, B., Genetic polymorphism in cytochrome P450 2D6 (CYP2D6): Population distribution of CYP2D6 activity. J. Toxicol. Environ. Health. B. Crit. Rev., 2009. **12**(5-6): p. 334-61.

87. Ingelman-Sundberg, M. and Rodriguez-Antona, C., Pharmacogenetics of drug-metabolizing enzymes: implications for a safer and more effective drug therapy. Philos. Trans. R. Soc. Lond. B. Biol. Sci., 2005. **360**(1460): p. 1563-70.

88. Lin, J.H., CYP induction-mediated drug interactions: in vitro assessment and clinical implications. Pharm. Res., 2006. **23**(6): p. 1089-116.

89. Court, M.H. and Greenblatt, D.J., Molecular genetic basis for deficient acetaminophen glucuronidation by cats: UGT1A6 is a pseudogene, and evidence for reduced diversity of expressed hepatic UGT1A isoforms. Pharmacogenetics, 2000. **10**(4): p. 355-69.

90. Giacomini, K.M., Huang, S.M., Tweedie, D.J., Benet, L.Z., Brouwer, K.L., Chu, X., Dahlin, A., Evers, R., Fischer, V., Hillgren, K.M., Hoffmaster, K.A., Ishikawa, T., Keppler, D., Kim, R.B., Lee, C.A., Niemi, M., Polli, J.W., Sugiyama, Y., Swaan, P.W., Ware, J.A., Wright, S.H., Yee, S.W., Zamek-Gliszczynski, M.J., and Zhang, L., Membrane transporters in drug development. Nat. Rev. Drug. Discov., 2010. **9**(3): p. 215-36.

91. Brandsch, M., Knutter, I., and Bosse-Doenecke, E., Pharmaceutical and pharmacological importance of peptide transporters. J. Pharm. Pharmacol., 2008. **60**(5): p. 543-85.

92. Meredith, D. and Price, R.A., Molecular modeling of PepT1-towards a structure. J. Membr. Biol., 2006. **213**(2): p. 79-88.

93. Brodin, B., Nielsen, C.U., Steffansen, B., and Frokjaer, S., Transport of peptidomimetic drugs by the intestinal di/tri-peptide transporter, PepT1. Pharmacol. Toxicol., 2002. **90**(6): p. 285-96.

94. 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.

95. Zhou, S.F. and Lai, X., An update on clinical drug interactions with the herbal antidepressant St. John's wort. Curr. Drug Metab., 2008. **9**(5): p. 394-409.

96. Juliano, R.L. and Ling, V., A surface glycoprotein modulating drug permeability in Chinese hamster ovary cell mutants. Biochim. Biophys. Acta, 1976. **455**(1): p. 152-62.

97. Chinn, L.W. and Kroetz, D.L., ABCB1 pharmacogenetics: progress, pitfalls, and promise. Clin. Pharmacol. Ther., 2007. **81**(2): p. 265-9.

98. Schinkel, A.H., P-Glycoprotein, a gatekeeper in the blood-brain barrier. Adv. Drug Deliv. Rev., 1999. **36**(2-3): p. 179-194.

99. Thiebaut, F., Tsuruo, T., Hamada, H., Gottesman, M.M., Pastan, I., and Willingham, M.C., Cellular localization of the multidrug-resistance gene product P-glycoprotein in normal human tissues. Proc. Natl. Acad. Sci. USA, 1987. **84**(21): p. 7735-8.

100. Greiner, B., Eichelbaum, M., Fritz, P., Kreichgauer, H.P., von Richter, O., Zundler, J., and Kroemer, H.K., The role of intestinal P-glycoprotein in the interaction of digoxin and rifampin. J. Clin. Invest., 1999. **104**(2): p. 147-53.

101. Fromm, M.F., Kim, R.B., Stein, C.M., Wilkinson, G.R., and Roden, D.M., Inhibition of P-glycoprotein-mediated drug transport: A unifying mechanism to explain the interaction between digoxin and quinidine. Circulation, 1999. **99**(4): p. 552-7.

102. Schinkel, A.H., Mayer, U., Wagenaar, E., Mol, C.A., van Deemter, L., Smit, J.J., van der Valk, M.A., Voordouw, A.C., Spits, H., van Tellingen, O., Zijlmans, J.M., Fibbe, W.E., and Borst, P., Normal viability and altered pharmacokinetics in mice lacking mdr1-type (drug-transporting) P-glycoproteins. Proc. Natl. Acad. Sci. USA, 1997. **94**(8): p. 4028-33.

103. Schinkel, A.H., Smit, J.J., van Tellingen, O., Beijnen, J.H., Wagenaar, E., van Deemter, L., Mol, C.A., van der Valk, M.A., Robanus-Maandag, E.C., te Riele, H.P., and et al., Disruption of the mouse mdr1a P-glycoprotein gene leads to a deficiency in the blood-brain barrier and to increased sensitivity to drugs. Cell, 1994. **77**(4): p. 491-502.

104. Edwards, J.E., Alcorn, J., Savolainen, J., Anderson, B.D., and McNamara, P.J., Role of P-glycoprotein in distribution of nelfinavir across the blood-mammary tissue barrier and blood-brain barrier. Antimicrob. Agents Chemother., 2005. **49**(4): p. 1626-8.

105. Alcorn, J., Lu, X., Moscow, J.A., and McNamara, P.J., Transporter gene expression in lactating and nonlactating human mammary epithelial cells using real-time reverse transcription-polymerase chain reaction. J. Pharmacol. Exp. Ther., 2002. **303**(2): p. 487-96.

106. Borowski, E., Bontemps-Gracz, M.M., and Piwkowska, A., Strategies for overcoming ABC-transporters-mediated multidrug resistance (MDR) of tumor cells. Acta Biochim. Pol., 2005. **52**(3): p. 609-27.

107. Fox, E. and Bates, S.E., Tariquidar (XR9576): a P-glycoprotein drug efflux pump inhibitor. Expert Rev. Anticancer Ther., 2007. **7**(4): p. 447-59.

108. Akhtar, N., Ahad, A., Khar, R.K., Jaggi, M., Aqil, M., Iqbal, Z., Ahmad, F.J., and Talegaonkar, S., The emerging role of P-glycoprotein inhibitors in drug delivery: a patent review. Expert Opin. Ther. Pat., 2011. **21**(4): p. 561-76.

109. Cole, S.P., Bhardwaj, G., Gerlach, J.H., Mackie, J.E., Grant, C.E., Almquist, K.C., Stewart, A.J., Kurz, E.U., Duncan, A.M., and Deeley, R.G., Overexpression of a transporter gene in a multidrug-resistant human lung cancer cell line. Science, 1992. **258**(5088): p. 1650-4.

110. Wakabayashi, K., Tamura, A., Saito, H., Onishi, Y., and Ishikawa, T., Human ABC transporter ABCG2 in xenobiotic protection and redox biology. Drug Metab. Rev., 2006. **38**(3): p. 371-91.

111. Robey, R.W., To, K.K., Polgar, O., Dohse, M., Fetsch, P., Dean, M., and Bates, S.E., ABCG2: a perspective. Adv Drug Deliv Rev, 2009. **61**(1): p. 3-13.

112. Zaher, H., Khan, A.A., Palandra, J., Brayman, T.G., Yu, L., and Ware, J.A., Breast cancer resistance protein (Bcrp/abcg2) is a major determinant of sulfasalazine absorption and elimination in the mouse. Mol. Pharm., 2006. **3**(1): p. 55-61.

113. Robey, R.W., Polgar, O., Deeken, J., To, K.W., and Bates, S.E., ABCG2: determining its relevance in clinical drug resistance. Cancer Metastasis Rev., 2007. **26**(1): p. 39-57.

114. Slovak, M.L., Ho, J.P., Bhardwaj, G., Kurz, E.U., Deeley, R.G., and Cole, S.P., Localization of a novel multidrug resistance-associated gene in the HT1080/DR4 and H69AR human tumor cell lines. Cancer Res., 1993. **53**(14): p. 3221-5.

115. Flens, M.J., Zaman, G.J., van der Valk, P., Izquierdo, M.A., Schroeijers, A.B., Scheffer, G.L., van der Groep, P., de Haas, M., Meijer, C.J., and Scheper, R.J., Tissue distribution of the multidrug resistance protein. Am. J. Pathol., 1996. **148**(4): p. 1237-47.

116. Kool, M., de Haas, M., Scheffer, G.L., Scheper, R.J., van Eijk, M.J., Juijn, J.A., Baas, F., and Borst, P., Analysis of expression of cMOAT (MRP2), MRP3, MRP4, and MRP5, homologues of the multidrug resistance-associated protein gene (MRP1), in human cancer cell lines. Cancer Res., 1997. **57**(16): p. 3537-47.

117. Keppler, D., Multidrug resistance proteins (MRPs, ABCCs): importance for pathophysiology and drug therapy. Handb Exp Pharmacol, 2011(201): p. 299-323.

118. Elferink, R.P., Tytgat, G.N., and Groen, A.K., Hepatic canalicular membrane 1: The role of mdr2 P-glycoprotein in hepatobiliary lipid transport. FASEB J., 1997. **11**(1): p. 19-28.

119. Eshkoli, T., Sheiner, E., Ben-Zvi, Z., and Holcberg, G., Drug transport across the placenta. Curr. Pharm. Biotechnol., 2011. **12**(5): p. 707-14.

120. Cui, Y.J., Cheng, X., Weaver, Y.M., and Klaassen, C.D., Tissue distribution, gender-divergent expression, ontogeny, and chemical induction of multidrug resistance transporter genes (Mdr1a, Mdr1b, Mdr2) in mice. Drug Metab. Dispos., 2009. **37**(1): p. 203-10.

121. Roch-Ramel, F., Renal transport of organic anions. Curr. Opin. Nephrol. Hypertens., 1998. **7**(5): p. 517-24.

122. Kusuhara, H. and Sugiyama, Y., Role of transporters in the tissue-selective distribution and elimination of drugs: transporters in the liver, small intestine, brain and kidney. J. Control. Release, 2002. **78**(1-3): p. 43-54.

123. Catania, V.A., Sanchez Pozzi, E.J., Luquita, M.G., Ruiz, M.L., Villanueva, S.S., Jones, B., and Mottino, A.D., Co-regulation of expression of phase II metabolizing enzymes and multidrug resistance-associated protein 2. Ann. Hepatol., 2004. **3**(1): p. 11-7.

124. Cui, Y., Konig, J., and Keppler, D., Vectorial transport by double-transfected cells expressing the human uptake transporter SLC21A8 and the apical export pump ABCC2. Mol. Pharmacol., 2001. **60**(5): p. 934-43.

125. Letschert, K., Komatsu, M., Hummel-Eisenbeiss, J., and Keppler, D., Vectorial transport of the peptide CCK-8 by double-transfected MDCKII cells stably expressing the organic anion transporter OATP1B3 (OATP8) and the export pump ABCC2. J. Pharmacol. Exp. Ther., 2005. **313**(2): p. 549-56.

126. Löscher, W. and Potschka, H., Role of drug efflux transporters in the brain for drug disposition and treatment of brain diseases. Prog. Neurobiol., 2005. **76**(1): p. 22-76.

127. Potschka, H., Fedrowitz, M., and Löscher, W., Multidrug resistance protein MRP2 contributes to blood-brain barrier function and restricts antiepileptic drug activity. J. Pharmacol. Exp. Ther., 2003. **306**(1): p. 124-31.

128. Jedlitschky, G., Hoffmann, U., and Kroemer, H.K., Structure and function of the MRP2 (ABCC2) protein and its role in drug disposition. Expert Opin Drug Metab Toxicol, 2006. **2**(3): p. 351-66.

129. Gottesman, M.M. and Ambudkar, S.V., Overview: ABC transporters and human disease. J. Bioenerg. Biomembr., 2001. **33**(6): p. 453-8.

130. König, J., Rost, D., Cui, Y., and Keppler, D., Characterization of the human multidrug resistance protein isoform MRP3 localized to the basolateral hepatocyte membrane. Hepatology, 1999. **29**(4): p. 1156-63.

131. Jedlitschky, G., Hoffmann, U., and Kroemer, H.K., Structure and function of the MRP2 (ABCC2) protein and its role in drug disposition. Expert Opin. Drug Metab. Toxicol., 2006. **2**(3): p. 351-66.

132. König, S.K., Herzog, M., Theile, D., Zembruski, N., Haefeli, W.E., and Weiss, J., Impact of drug transporters on cellular resistance towards saquinavir and darunavir. J. Antimicrob. Chemother., 2010. **65**(11): p. 2319-28.

133. Zelcer, N., Huisman, M.T., Reid, G., Wielinga, P., Breedveld, P., Kuil, A., Knipscheer, P., Schellens, J.H., Schinkel, A.H., and Borst, P., Evidence for two interacting ligand binding sites in human multidrug resistance protein 2 (ATP binding cassette C2). J. Biol. Chem., 2003. **278**(26): p. 23538-44.

134. Evers, R., de Haas, M., Sparidans, R., Beijnen, J., Wielinga, P.R., Lankelma, J., and Borst, P., Vinblastine and sulfinpyrazone export by the multidrug resistance protein MRP2 is associated with glutathione export. Br. J. Cancer, 2000. **83**(3): p. 375-83.

135. Van Aubel, R.A., Koenderink, J.B., Peters, J.G., Van Os, C.H., and Russel, F.G., Mechanisms and interaction of vinblastine and reduced glutathione transport in membrane vesicles by the rabbit multidrug resistance protein Mrp2 expressed in insect cells. Mol. Pharmacol., 1999. **56**(4): p. 714-9.

136. MacDougall, C. and Guglielmo, B.J., Pharmacokinetics of valaciclovir. J. Antimicrob. Chemother., 2004. **53**(6): p. 899-901.

137. Sugawara, M., Huang, W., Fei, Y.J., Leibach, F.H., Ganapathy, V., and Ganapathy, M.E., Transport of valganciclovir, a ganciclovir prodrug, via peptide transporters PEPT1 and PEPT2. J. Pharm. Sci., 2000. **89**(6): p. 781-9.

138. Cui, Y., Konig, J., Leier, I., Buchholz, U., and Keppler, D., Hepatic uptake of bilirubin and its conjugates by the human organic anion transporter SLC21A6. J Biol Chem, 2001. **276**(13): p. 9626-30.

139. Abe, T., Kakyo, M., Tokui, T., Nakagomi, R., Nishio, T., Nakai, D., Nomura, H., Unno, M., Suzuki, M., Naitoh, T., Matsuno, S., and Yawo, H., Identification of a novel gene family encoding human liver-specific organic anion transporter LST-1. J. Biol. Chem., 1999. **274**(24): p. 17159-63.

140. König, J., Cui, Y., Nies, A.T., and Keppler, D., A novel human organic anion transporting polypeptide localized to the basolateral hepatocyte membrane. Am. J. Physiol. Gastrointest. Liver Physiol., 2000. **278**(1): p. G156-64.

141. Kalliokoski, A. and Niemi, M., Impact of OATP transporters on pharmacokinetics. Br. J. Pharmacol., 2009. **158**(3): p. 693-705.

142. Smith, D.A., van de Waterbeemd, H., and Walke, D.K., Pharmacokinetics and metabolism in drug design. Methods and principles in medicinal chemistry. Second ed. Vol. 31. 2006, Weinheim, Germany.: Wiley-VCH.

143. Neuvonen, P.J., Niemi, M., and Backman, J.T., Drug interactions with lipid-lowering drugs: mechanisms and clinical relevance. Clin Pharmacol Ther, 2006. **80**(6): p. 565-81.

144. Link, E., Parish, S., Armitage, J., Bowman, L., Heath, S., Matsuda, F., Gut, I., Lathrop, M., and Collins, R., SLCO1B1 variants and statin-induced myopathy-a genomewide study. N. Engl. J. Med., 2008. **359**(8): p. 789-99.

145. Niemi, M., Pasanen, M.K., and Neuvonen, P.J., SLCO1B1 polymorphism and sex affect the pharmacokinetics of pravastatin but not fluvastatin. Clin. Pharmacol. Ther., 2006. **80**(4): p. 356-66.

146. Kullak-Ublick, G.A., Fisch, T., Oswald, M., Hagenbuch, B., Meier, P.J., Beuers, U., and Paumgartner, G., Dehydroepiandrosterone sulfate (DHEAS): identification of a carrier protein in human liver and brain. FEBS Lett., 1998. **424**(3): p. 173-6.

147. Dresser, G.K., Kim, R.B., and Bailey, D.G., Effect of grapefruit juice volume on the reduction of fexofenadine bioavailability: possible role of organic anion transporting polypeptides. Clin. Pharmacol. Ther., 2005. **77**(3): p. 170-7.

148. Dresser, G.K., Bailey, D.G., Leake, B.F., Schwarz, U.I., Dawson, P.A., Freeman, D.J., and Kim, R.B., Fruit juices inhibit organic anion transporting polypeptide-mediated drug uptake to decrease the oral availability of fexofenadine. Clin. Pharmacol. Ther., 2002. **71**(1): p. 11-20.

149. Cros, Action de l' 'alcohol amylique sur l' 'organisme" , in Faculty of Medicine. 1863, Universitée de Strasbourg: Strasbourg.

150. Crum-Brown, A. and Fraser, T.R., On the connection between chemical constitution and physiological action; With special reference to the physiological action of the salts of the ammonium bases derived from strychnia, brucia, thebata, codeia, morphia, and nicotia. Trans. Roy. Soc. Edinburgh, 1868-1869. **25**: p. 1-53.

151. Körner, W., Fatti per servire alla determinazione del luogo chimico nelle sostanze aromatiche. Giornale di Scienze Naturali ed Economiche, 1869. **5**: p. 212-256.

152. Körner, W., Studi sulla Isomeria delle Cosi̊ Dette Sostanze Aromatiche a Sei Atomi di Carbonio. Gazz. Chim. It., 1874. **4**: p. 242.

153. Mills, E.J., On melting point and boiling point as related to composition. Philos. Mag., 1884. **17**: p. 127-187.

154. Richet, M.C., Noté sur la Rapport entre la Toxicité et les Propriétés Physiques des Corps. Compt. Rend. Soc. Biol., 1893. **45**: p. 775-776.

155. Meyer, K.H., Contribution to the theory of narcosis. Trans. Faraday Soc., 1937(33): p. 1060-1068.

156. Overton, C.E., Studien über die Narkose, zugleich ein Beitrag zur allgemeinen Pharmakologie. 1901, Gustav Fischer: Jena.

157. Traube, I., Theorie der Osmose und Narkose. Arch. für die ges. Physiol., 1904. **105**: p. 541-558.

158. Hammett, L.P., Linear free energy relationships in rate and equilibrium phenomena. Trans. Faraday Soc., 1938. **34**: p. 156-165.

159. Wiener, H., Correlation of heat of isomerization, and differences in heats of vaporization of isomers, among the paraffin hydrocarbons. J. Am. Chem. Soc., 1947a. **69**: p. 1636-1638.

160. Wiener, H., Influence of interatomic forces on paraffin properties. J. Chem. Phys., 1947b. **15**: p. 766.

161. Wiener, H., Structural determination of paraffin boiling points. J. Am. Chem. Soc., 1947c. **69**: p. 17-20.

162. Platt, J.R., Influence of neighbor bonds on additive bond properties in paraffins. J. Chem. Phys., 1947. **15**: p. 419-420.

163. Hansch, C., Maloney, P.P., Fujita, T., and Muir, R.M., Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. Nature, 1962. **194**: p. 178-180.

164. Free, S.M. and Wilson, J.W., A mathematical contribution to structure–activity studies. J. Med. Chem., 1964. **7**: p. 395-399.

165. Stanton, D.T. and Jurs, P.C., Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure-property relationship studies. Anal. Chem., 1990. **62**: p. 2323-2329.

166. Katritzky, A.R., Mu, L., Lobanov, V.S., and Karelson, M., Correlation of boiling points with molecular structure. 1. A training set of 298 diverse organics and a test set of 9 simple inorganics. J. Phys. Chem., 1996. **100**: p. 10400-10407.

167. Sheridan, R.P. and Kearsley, S.K., Why do we need so many chemical similarity search methods? Drug Discov. Today, 2002. **7**(17): p. 903-11.

168. McGaughey, G.B., Sheridan, R.P., Bayly, C.I., Culberson, J.C., Kreatsoulas, C., Lindsley, S., Maiorov, V., Truchon, J.F., and Cornell, W.D., Comparison of topological, shape, and docking methods in virtual screening. J. Chem. Inf. Model., 2007. **47**(4): p. 1504-19.

169. Bonchev, D. and Rouvray, D.H., Chemical graph theory: Introduction and fundamentals. Mathematical Chemistry Series, ed. D.H. Rouvray. Vol. 1. 1991, New York: Abacus.

170. Weininger, D., SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J. Chem. Inf. Model., 1988. **28**(31).

171. Weininger, D., Weininger, A., and Weininger, J.L., SMILES. 2. Algorithm for generation of unique SMILES notation. J. Chem. Inf. Model., 1989. **29**(2).

172. Weininger, D. Daylight SMARTS Theory Manual. [cited 2011 6. 9. 2011]; Available from: http://www.daylight.com/dayhtml_tutorials/languages/smarts/index.html#INTRO.

173. Hassan, M., Brown, R.D., Varma-O'Brien, S., and Rogers, D., Cheminformatics analysis and learning in a data pipelining environment. Mol. Divers., 2006. **10**(3): p. 283-99.

174. Durant, J.L., Leland, B.A., Henry, D.R., and Nourse, J.G., Reoptimization of MDL keys for use in drug discovery. J. Chem. Inf. Comput. Sci., 2002. **42**(6): p. 1273-80.

175. Todeschini, R. and Consonni, V., Molecular descriptors for chemoinformatics. Methods and principles in medicinal chemistry ed. R. Mannhold, Kubinyi, H., and Folkers, G. 2000, Weinheim, New York: Wiley-VCH.

176. Bakken, G. and Jurs, P.C., Prediction of Hydroxyl Radical Rate Constants from Molecular Structure. J. Chem. Inf. Comput. Sci., 1999. **39**: p. 1064-75.

177. Hammann, F., Gutmann, H., Baumann, U., Helma, C., and Drewe, J., Classification of cytochrome p(450) activities using machine learning methods. Mol. Pharm., 2009. **6**(6): p. 1920-6.

178. Veber, D.F., Johnson, S.R., Cheng, H.Y., Smith, B.R., Ward, K.W., and Kopple, K.D., Molecular properties that influence the oral bioavailability of drug candidates. J Med Chem, 2002. **45**(12): p. 2615-23.

179. Iyer, M., Tseng, Y.J., Senese, C.L., Liu, J., and Hopfinger, A.J., Prediction and mechanistic interpretation of human oral drug absorption using MI-QSAR analysis. Mol Pharm, 2007. **4**(2): p. 218-31.

180. Moriguchi, I., Hirono, S., Liu, Q., and Nakagome, I., Simple method of calculating octanol/water partition coefficient. Chem. Pharm. Bull., 1992. **40**: p. 127-130.

181. Viswanadhan, V.N., Reddy, M.R., Bacquet, R.J., and Erion, M.D., Assessment of methods used for predicting lipophilicity: application to nucleosides and nucleoside bases. . J. Comput. Chem., 1993. **14**: p. 1-4.

182. Ghose, A.K., Viswanadhan, V.N., and Wendoloski, J.J., Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: an analysis of ALOGP and CLOGP methods. . J. Phys. Chem. A., 1998. **102**: p. 2762-2772.

183. Leo, A., Jow, P.Y.C., Silipo, C., and Hansch, C., Calculation of hydrophobic constant (log P) from p and f constants. . J.Med.Chem., 1975. **18**: p. 865-868.

184. Chou, J.T. and Jurs, P.C., Computer-assisted computation of partial coefficients from molecular structures using fragment constants. J. Chem. Inf. Comput. Sci., 1979. **19**: p. 172-178.

185. Jeffrey, G.A., An Introduction to Hydrogen Bonding, in Topics in Physical Chemistry. 1997, Oxford University Press.

186. Habgood, M.D., Begley, D.J., and Abbott, N.J., Determinants of passive drug entry into the central nervous system. Cell. Mol. Neurobiol., 2000. **20**(2): p. 231-53.

187. Gratton, J.A., Abraham, M.H., Bradbury, M.W., and Chadha, H.S., Molecular factors influencing drug transfer across the blood-brain barrier. J Pharm Pharmacol, 1997. **49**(12): p. 1211-6.

188. Pearlman, R.S., Molecular Surface Area and Volumes and their Use in Structure/Activity Relationships, in Physical Chemical Properties of Drugs. 1980, Marcel Drekker, Inc.: New York.

189. Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E., and Willighagen, E., 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.

190. Ertl, P., Rohde, B., and Selzer, P., 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.

191. Randic, M., Generalized molecular descriptors. J. Math. Chem., 1991. **7**: p. 155-168.

192. Randic, M., The nature of the chemical structure. J. Math. Chem., 1990. **4**: p. 157-184.

193. Randic, M., Novel graph theoretical approach to heteroatoms in quantitative structure–activity relationships. . Chemom. Intell. Lab. Syst., 1991. **10**: p. 213-227.

194. Burden, F.R., Molecular identification number for substructure searches. J. Chem. Inf. Comput., 1989. **29**: p. 225-227.

195. Pearlman, R.S. and Smith, K.M., Metric Validation and the Receptor-Relevant Subspace Concept. J. Chem. Inf. Comput. Sci., 1999. **39**: p. 28-35.

196. Kier, L.B., A shape index from molecular graphs. Quant. Struct. -Act. Relat., 1985. **4**: p. 109-116.

197. Wessel, M.D., Jurs, P.C., Tolan, J.W., and Muskal, S.M., Prediction of human intestinal absorption of drug compounds from molecular structure. J. Chem. Inf. Comput. Sci., 1998. **38**: p. 726-735.

198. Tanford, C., Physical Chemistry of Macromolecules. 1961, New York: John Wiley & Sons, Inc.

199. 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**(331): p. 331-337.

200. Bath, P.A., Poirrette, A.R., Willett, P., and Allen, F.H., The extent of the relationship between the graph-theoretical and the geometrical shape coefficients of chemical compounds. J.Chem.Inf.Comput.Sci., 1995. **4**(35): p. 714-716.

201. Fisher, R.A., The use of multiple measurements in taxonomic problems. Annals of Eugenics, 1936. **7**(2): p. 179-188.

202. Breiman, L., Bagging predictors. Machine Learning, 1996. **24**: p. 123-140.

203. Kim, S.H., An Extension of CART's Pruning Algorithm, in Program Statistics Research Technical Report No. 91-11. 1991.

204. Breiman, L., Classification and regression trees. 1st ed. 1984, Boca Raton: Chapman & Hall/CRC.

205. Gini, C., Variabitlità e mutabilità. Memorie di metodologica statistica, 1912.

206. Sonquist, J.A. and Morgan, J.N., The detection of interaction effects. 1964, Survey research center, University of Michigan: Ann Arbor. p. 296.

207. Breiman, L., Random forests. Machine Learning, 2001. **45**(1): p. 5 - 32.

208. Breiman, L., Bagging predictors. Machine Learning, 1996. **24**(2): p. 123-140.

209. Segal, M.R., Machine learning benchmarks and Random Forest Regression, U.o.C. Centre of Bioinformatics & Molecular Biostatistics, Editor. 2004: San Francisco.

210. McCulloch, W. and Pitts, W.A., A logical calculus of ideas immanent in nervous activity. Bull. Math. Biophys., 1943. **5**: p. 115 - 33.

211. Cortes, C. and Vapnik, V., Support-vector networks. Machine Learning, 1995. **20**(3): p. 273-297.

212. Aizerman, M., Braverman, E., and Rozonoer, L., Theoretical foundations of the potential function method in pattern recognition. Autom. Remote Control, 1964. **25**: p. 821-837.

213. Unwin, S., The probability of god: a simple calculation that proves the ultimate truth., ed. C.a. Formun. 2003, New York.

214. Smith, A.F.M. and Roberts, G.O., Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. J. Roy. Stat. Soc. B, 1993. **55**(1): p. 3-23.

215. Tierney, L., Markov chains for exploring posterior distributions. Ann. Stat., 1994. **22**(4): p. 1701-62.

216. Langley, P., Iba, W., and Thomas, K. An analysis of Bayesian Classifiers. in Tenth National Conference on Artificial Intelligence. 1992: AAAI Press.

217. Kononenko, I., Comparison of inductive and naive Bayesian learning approaches to automatic knowledge acquisition, in Current trends in knowledge acquisition, B. Wielinga, Editor. 1990, OS Press.

218. Zhang, H., The Optimality of Naive Bayes. Proceedings of the Seventeenth Florida Artificial Intelligence Research Society Conference, 2004: p. 562-567.

219. Domingos, P. and Pazzani, M., On the optimality of the simple Bayesian classifier under zero-one loss Machine Learning, 1997. **29**(2-3): p. 103-130.

220. Friedman, J.H., On bias, variance, 0/1-loss, and the curse-of-dimensionality. Data Mining and Knowledge Discovery 1996. **1**: p. 55-77.

221. Russel, S. and Norvig, P., Artificial intelligence: A modern approach. 2nd ed. 2002, Upper Saddle River, NJ: Prentice Hall.

222. Helma, C., Lazy structure-activity relationships (lazar) for the prediction of rodent carcinogenicity and Salmonella mutagenicity. Mol. Divers., 2006. **10**(2): p. 147-58.

223. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutmann, P., and Witten, I.H., The WEKA data mining software: An update. SIGKDD Explorations, 2009. **11**(1): p. 10-18.

224. Bonabeau, E., Dorigo, M., and Theraulaz, G., Inspiration for optimization from social insect behaviour. Nature, 2000. **406**(6791): p. 39-42.

225. Gross, S., Aron, S., Deneubourg, J.L., and Pasteels, J.M., Self-organized shortcuts in the argentine ant. Naturwissenschaften, 1989. **76**: p. 579-81.

226. Bonabeau, E., Theraulaz, G., Deneubourg, J.L., Aron, S., and Camazine, S., Self-organization in social insects. Trends Ecol. Evol., 1997. **12**(5): p. 188-93.

227. Youden, W.J., Index for rating diagnostic tests. Cancer, 1950. **3**(1): p. 32-35.

228. Kohavi, R.A., A study of cross-validation and bootstrap for accuracy estimation and model selection. Proc. 14th Int. Jt. Conf. Artif. Intell., 1995. **2**(12): p. 1137-1143.

229. Hassinen, T. and Peräkylä, M., New energy terms for reduced protein models implemented in an off-lattice force field. J. Comput. Chem., 2001. **22**: p. 1229-1242.

230. Dollery, C., Therapeutic Drugs. 2th ed, ed. C. Dollery. Vol. 1 and 2 1999, Edinburgh: Churchill Livingstone.

231. Cooper, G.F. and Herskovits, E., A Bayesian method for the induction of probabilistic networks from data. Machine Learning, 1992. **9**(4): p. 309-347.

232. Wang, Z., Yan, A., Yuan, Q., and Gasteiger, J., Explorations into modeling human oral bioavailability. Eur. J. Med. Chem., 2008. **43**(11): p. 2442-52.

233. Hammann, F., Gutmann, H., Vogt, N., Helma, C., and Drewe, J., Prediction of adverse drug reactions using decision tree modeling. Clin Pharmacol Ther, 2010. **88**(1): p. 52-9.

234. Palm, K., Luthman, K., Ungell, A.L., Strandlund, G., and Artursson, P., Correlation of drug absorption with molecular surface properties. J. Pharm. Sci., 1996. **85**(1): p. 32-9.

235. Winiwarter, S., Bonham, N.M., Ax, F., Hallberg, A., Lennernas, H., and Karlen, A., 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.

236. Palm, K., Stenberg, P., Luthman, K., and Artursson, P., Polar molecular surface properties predict the intestinal absorption of drugs in humans. Pharm. Res., 1997. **14**(5): p. 568-71.

237. Grass, G.M. and Sinko, P.J., Effect of diverse datasets on the predictive capability of ADME models in drug discovery. Drug Discov. Today 2001. **6**: p. 54-61.

238. Hou, T., Wang, J., Zhang, W., and Xu, X., ADME evaluation in drug discovery. 7. Prediction of oral absorption by correlation and classification. J Chem Inf Model, 2007. **47**(1): p. 208-18.

239. Linnankoski, J., Makela, J.M., Ranta, V.P., Urtti, A., and Yliperttula, M., Computational prediction of oral drug absorption based on absorption rate constants in humans. J. Med. Chem., 2006. **49**(12): p. 3674-81.

240. Zhao, Y.H., Abraham, M.H., Le, J., Hersey, A., Luscombe, C.N., Beck, G., Sherborne, B., and Cooper, I., Rate-limited steps of human oral absorption and QSAR studies. Pharm. Res., 2002. **19**(10): p. 1446-57.

241. Pearlman, R.S. and Smith, K.M., Metric Validation and the Receptor-Relevant Subspace Concept. J. Chem. Inf. Comput. Sci., 1999. **39**(1): p. 28-35.

242. Pearlman, R.S. and Smith, K.M., Metric Validation and the Receptor Relevant Subspace Concept. J. Chem. Inf. Comput .Sci., 1999. **39**: p. 28-35.

243. Manallack, D.T., Tehan, B.G., Gancia, E., Hudson, B.D., Ford, M.G., Livingstone, D.J., Whitley, D.C., and Pitt, W.R., A consensus neural network-based technique for discriminating soluble and poorly soluble compounds. J. Chem. Inf. Comput. Sci., 2003. **43**(2): p. 674-9.

244. Obrezanova, O. and Segall, M.D., Gaussian processes for classification: QSAR modeling of ADMET and target activity. J. Chem. Inf. Model., 2010. **50**(6): p. 1053-61.

245. Shen, J., Cheng, F., Xu, Y., Li, W., and Tang, Y., Estimation of ADME properties with substructure pattern recognition. J. Chem. Inf. Model., 2010. **50**(6): p. 1034-41.

246. Hou, T., Wang, J., and Li, Y., ADME evaluation in drug discovery. 8. The prediction of human intestinal absorption by a support vector machine. J. Chem. Inf. Model., 2007. **47**(6): p. 2408-15.

247. Ertl, P., Rohde, B., and Selzer, P., 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.

248. Zhao, Y.H., Le, J., Abraham, M.H., Hersey, A., Eddershaw, P.J., Luscombe, C.N., Butina, D., Beck, G., Sherborne, B., Cooper, I., and Platts, J.A., Evaluation of human intestinal absorption data and subsequent derivation of a quantitative structure-activity relationship (QSAR) with the Abraham descriptors. J. Pharm. Sci., 2001. **90**(6): p. 749-84.

249. Niwa, T., Using general regression and probabilistic neural networks to predict human intestinal absorption with topological descriptors derived from two-dimensional chemical structures. J. Chem. Inf. Comput. Sci., 2003. **43**(1): p. 113-9.

250. Bai, J.P., Utis, A., Crippen, G., He, H.D., Fischer, V., Tullman, R., Yin, H.Q., Hsu, C.P., Jiang, L., and Hwang, K.K., Use of classification regression tree in predicting oral absorption in humans. J. Chem. Inf. Comput. Sci., 2004. **44**(6): p. 2061-9.

251. Liu, H.X., Hu, R.J., Zhang, R.S., Yao, X.J., Liu, M.C., Hu, Z.D., and Fan, B.T., The prediction of human oral absorption for diffusion rate-limited drugs based on heuristic method and support vector machine. J. Comput. Aided. Mol. Des., 2005. **19**(1): p. 33-46.

252. Jones, R., Connolly, P.C., Klamt, A., and Diedenhofen, M., Use of surface charges from DFT calculations to predict intestinal absorption. J. Chem. Inf. Model., 2005. **45**(5): p. 1337-42.

253. Deconinck, E., Hancock, T., Coomans, D., Massart, D.L., and Heyden, Y.V., Classification of drugs in absorption classes using the classification and regression trees (CART) methodology. J. Pharm. Biomed. Anal., 2005. **39**(1-2): p. 91-103.

254. Iyer, M., Tseng, Y.J., Senese, C.L., Liu, J., and Hopfinger, A.J., Prediction and mechanistic interpretation of human oral drug absorption using MI-QSAR analysis. Mol. Pharm., 2007. **4**(2): p. 218-31.

255. Yan, A., Wang, Z., and Cai, Z., Prediction of human intestinal absorption by GA feature selection and support vector machine regression. Int. J. Mol. Sci., 2008. **9**(10): p. 1961-76.

256. Reynolds, D.P., Lanevskij, K., Japertas, P., Didziapetris, R., and Petrauskas, A., Ionization-specific analysis of human intestinal absorption. J Pharm Sci, 2009. **98**(11): p. 4039-54.

257. Guerra, A., Campillo, N.E., and Paez, J.A., Neural computational prediction of oral drug absorption based on CODES 2D descriptors. Eur. J. Med. Chem., 2010. **45**(3): p. 930-40.

258. Hammann, F., Gutmann, H., Jecklin, U., Maunz, A., Helma, C., and Drewe, J., Development of decision tree models for substrates, inhibitors, and inducers of p-glycoprotein. Curr. Drug Metab., 2009. **10**(4): p. 339-46.

259. Hammann, F., Gutmann, H., Vogt, N., Helma, C., and Drewe, J., Prediction of adverse drug reactions using decision tree modeling. Clin. Pharmacol. Ther., 2010. **88**(1): p. 52-9.

260. Suenderhauf, C., Hammann, F., Maunz, A., Helma, C., and Huwyler, J., Combinatorial QSAR modeling of human intestinal absorption. Mol Pharm, 2011. **8**(1): p. 213-24.

261. Pardridge, W.M., Triguero, D., Yang, J., and Cancilla, P.A., Comparison of in vitro and in vivo models of drug transcytosis through the blood-brain barrier. J. Pharmacol. Exp. Ther., 1990. **253**(2): p. 884-91.

262. Smith, Q.R. and Takasato, Y., Kinetics of amino acid transport at the blood-brain barrier studied using an in situ brain perfusion technique. Ann. N. Y. Acad. Sci., 1986. **481**: p. 186-201.

263. Greig, N.H., Momma, S., Sweeney, D.J., Smith, Q.R., and Rapoport, S.I., Facilitated transport of melphalan at the rat blood-brain barrier by the large neutral amino acid carrier system. Cancer Res., 1987. **47**(6): p. 1571-6.

264. Momma, S., Aoyagi, M., Rapoport, S.I., and Smith, Q.R., Phenylalanine transport across the blood-brain barrier as studied with the in situ brain perfusion technique. J. Neurochem., 1987. **48**(4): p. 1291-300.

265. Levin, V.A., Relationship of octanol/water partition coefficient and molecular weight to rat brain capillary permeability. J. Med. Chem., 1980. **23**(6): p. 682-4.

266. Gratton, J.A., Abraham, M.H., Bradbury, M.W., and Chadha, H.S., Molecular factors influencing drug transfer across the blood-brain barrier. J. Pharm. Pharmacol., 1997. **49**(12): p. 1211-6.

267. Tamai, I., Yamashita, J., Kido, Y., Ohnari, A., Sai, Y., Shima, Y., Naruhashi, K., Koizumi, S., and Tsuji, A., Limited distribution of new quinolone antibacterial agents into brain caused by multiple efflux transporters at the blood-brain barrier. J. Pharmacol. Exp. Ther., 2000. **295**(1): p. 146-52.

268. Murakami, H., Takanaga, H., Matsuo, H., Ohtani, H., and Sawada, Y., Comparison of blood-brain barrier permeability in mice and rats using in situ brain perfusion technique. Am. J. Physiol. Heart Circ. Physiol., 2000. **279**(3): p. H1022-8.

269. Liu, X., Tu, M., Kelly, R.S., Chen, C., and Smith, B.J., Development of a computational approach to predict blood-brain barrier permeability. Drug Metab. Dispos., 2004. **32**(1): p. 132-9.

270. Youdim, K.A., Qaiser, M.Z., Begley, D.J., Rice-Evans, C.A., and Abbott, N.J., Flavonoid permeability across an in situ model of the blood-brain barrier. Free Radic. Biol. Med., 2004. **36**(5): p. 592-604.

271. Parepally, J.M., Mandula, H., and Smith, Q.R., Brain uptake of nonsteroidal anti-inflammatory drugs: ibuprofen, flurbiprofen, and indomethacin. Pharm. Res., 2006. **23**(5): p. 873-81.

272. Summerfield, S.G., Read, K., Begley, D.J., Obradovic, T., Hidalgo, I.J., Coggon, S., Lewis, A.V., Porter, R.A., and Jeffrey, P., Central nervous system drug disposition: the relationship between in situ brain permeability and brain free fraction. J. Pharmacol. Exp. Ther., 2007. **322**(1): p. 205-13.

273. Lanevskij, K., Japertas, P., Didziapetris, R., and Petrauskas, A., Ionization-specific prediction of blood-brain permeability. J. Pharm. Sci., 2009. **98**(1): p. 122-34.

274. Fischer, H., Gottschlich, R., and Seelig, A., Blood-brain barrier permeation: molecular parameters governing passive diffusion. J. Membr. Biol., 1998. **165**(3): p. 201-11.

275. Wang, X., Ratnaraj, N., and Patsalos, P.N., The pharmacokinetic inter-relationship of tiagabine in blood, cerebrospinal fluid and brain extracellular fluid (frontal cortex and hippocampus). Seizure, 2004. **13**(8): p. 574-81.

276. Schinkel, A.H., Wagenaar, E., Mol, C.A., and van Deemter, L., P-glycoprotein in the blood-brain barrier of mice influences the brain penetration and pharmacological activity of many drugs. J. Clin. Invest., 1996. **97**(11): p. 2517-24.

277. Goodwin, J.T. and Clark, D.E., In silico predictions of blood-brain barrier penetration: considerations to "keep in mind". J Pharmacol Exp Ther, 2005. **315**(2): p. 477-83.

278. Van de Waterbeemd, H., Camenisch, G., Folkers, G., Chretien, J.R., and Raevsky, O.A., Estimation of blood-brain barrier crossing of drugs using molecular size and shape, and H-bonding descriptors. J. Drug. Target., 1998. **6**(2): p. 151-65.

279. Waterhouse, R.N., Determination of lipophilicity and its use as a predictor of blood-brain barrier penetration of molecular imaging agents. Mol. Imaging Biol., 2003. **5**(6): p. 376-89.

280. Pardridge, W.M., The blood-brain barrier: bottleneck in brain drug development. NeuroRx, 2005. **2**(1): p. 3-14.

281. Wang, R.B., Kuo, C.L., Lien, L.L., and Lien, E.J., Structure-activity relationship: analyses of P-glycoprotein substrates and inhibitors. J. Clin. Pharm. Ther., 2003. **28**(3): p. 203-28.

282. Kelder, J., Grootenhuis, P.D., Bayada, D.M., Delbressine, L.P., and Ploemen, J.P., Polar molecular surface as a dominating determinant for oral absorption and brain penetration of drugs. Pharm. Res., 1999. **16**(10): p. 1514-9.

283. 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.

284. Huang, J., Ma, G., Muhammad, I., and Cheng, Y., Identifying P-glycoprotein substrates using a support vector machine optimized by a particle swarm. J. Chem. Inf. Model., 2007. **47**(4): p. 1638-47.

285. Didziapetris, R., Japertas, P., Avdeef, A., and Petrauskas, A., Classification analysis of P-glycoprotein substrate specificity. J. Drug. Target., 2003. **11**(7): p. 391-406.

286. Norinder, U. and Haeberlein, M., Computational approaches to the prediction of the blood-brain distribution. Adv. Drug Deliv. Rev., 2002. **54**(3): p. 291-313.

287. Abraham, M.H., The factors that influence permeation across the blood-brain barrier. Eur. J. Med. Chem., 2004. **39**(3): p. 235-40.

288. Klein, C.T., Kaiser, D., and Ecker, G., Topological distance based 3D descriptors for use in QSAR and diversity analysis. J. Chem. Inf. Comput. Sci., 2004. **44**(1): p. 200-9.

289. Schuur, J.H., Selzer, P., and Gasteiger, J., The Coding of the Three-Dimensional Structure of Molecules by Molecular Transforms and Its Application to Structure-Spectra Correlations and Studies of Biological Activity. Journal of Chemical Information and Computer Sciences, 1996. **36**(2): p. 334-344.

290. Strolin Benedetti, M., Frigerio, E., Tocchetti, P., Brianceschi, G., Castelli, M.G., Pellizzoni, C., and Dostert, P., Stereoselective and species-dependent kinetics of reboxetine in mouse and rat. Chirality, 1995. **7**(4): p. 285-9.

291. Pham, Y.T., Nosten, F., Farinotti, R., White, N.J., and Gimenez, F., Cerebral uptake of mefloquine enantiomers in fatal cerebral malaria. Int. J. Clin. Pharmacol. Ther., 1999. **37**(1): p. 58-61.

292. NCBI, N.C.f.B.I. PubChem. [cited 2011 06.09.2011]; Available from: http://pubchem.ncbi.nlm.nih.gov.

293. Willett, P., Similarity-based virtual screening using 2D fingerprints. Drug Discov. Today, 2006. **11**(23-24): p. 1046-53.

294. Meskin, M.S. and Lien, E.J., QSAR analysis of drug excretion into human breast milk. J. Clin. Hosp. Pharm., 1985. **10**(3): p. 269-78.

295. Agatonovic-Kustrin, S., Ling, L.H., Tham, S.Y., and Alany, R.G., Molecular descriptors that influence the amount of drugs transfer into human breast milk. J. Pharm. Biomed. Anal., 2002. **29**(1-2): p. 103-19.

296. Fleishaker, J.C., Desai, N., and McNamara, P.J., Factors affecting the milk-to-plasma drug concentration ratio in lactating women: physical interactions with protein and fat. J Pharm Sci, 1987. **76**(3): p. 189-93.

297. Jan H. Schuur, Paul Selzer, and Gasteiger, J., The Coding of the Three-Dimensional Structure of Molecules by Molecular Transforms and Its Application to Structure-Spectra Correlations and Studies of Biological Activity. Journal of Chemical Information and Computer Sciences, 1996. **36**(2): p. 334-344.

298. Consonni, V., Todeschini, R., and Pavan, M., Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors. J. Chem. Inf. Comput. Sci., 2002. **42**(3): p. 682-92.

299. Matsson, P., Englund, G., Ahlin, G., Bergstrom, C.A., Norinder, U., and Artursson, P., A global drug inhibition pattern for the human ATP-binding cassette transporter breast cancer resistance protein (ABCG2). J. Pharmacol. Exp. Ther., 2007. **323**(1): p. 19-30.

300. Alcorn, J., Lu, X., Moscow, J.A., and McNamara, P.J., Transporter gene expression in lactating and nonlactating human mammary epithelial cells using real-time reverse transcription-polymerase chain reaction. J Pharmacol Exp Ther, 2002. **303**(2): p. 487-96.

301. Gombar, V.K., Polli, J.W., Humphreys, J.E., Wring, S.A., and Serabjit-Singh, C.S., Predicting P-glycoprotein substrates by a quantitative structure-activity relationship model. J. Pharm. Sci., 2004. **93**(4): p. 957-68.

302. Nies, A.T., Schwab, M., and Keppler, D., Interplay of conjugating enzymes with OATP uptake transporters and ABCC/MRP efflux pumps in the elimination of drugs. Expert Opin. Drug Metab. Toxicol., 2008. **4**(5): p. 545-68.

303. Cassio, D., Macias, R.I., Grosse, B., Marin, J.J., and Monte, M.J., Expression, localization, and inducibility by bile acids of hepatobiliary transporters in the new polarized rat hepatic cell lines, Can 3-1 and Can 10. Cell Tissue Res., 2007. **330**(3): p. 447-60.

304. Sai, K., Saito, Y., Itoda, M., Fukushima-Uesaka, H., Nishimaki-Mogami, T., Ozawa, S., Maekawa, K., Kurose, K., Kaniwa, N., Kawamoto, M., Kamatani, N., Shirao, K., Hamaguchi, T., Yamamoto, N., Kunitoh, H., Ohe, Y., Yamada, Y., Tamura, T., Yoshida, T., Minami, H., Matsumura, Y., Ohtsu, A., Saijo, N., and Sawada, J., Genetic variations and haplotypes of ABCC2 encoding MRP2 in a Japanese population. Drug Metab. Pharmacokinet., 2008. **23**(2): p. 139-47.

305. Suzuki, H. and Sugiyama, Y., Single nucleotide polymorphisms in multidrug resistance associated protein 2 (MRP2/ABCC2): its impact on drug disposition. Adv. Drug Deliv. Rev., 2002. **54**(10): p. 1311-31.

306. Zhang, H., Xiang, M.L., Zhao, Y.L., Wei, Y.Q., and Yang, S.Y., Support vector machine and pharmacophore-based prediction models of multidrug-resistance protein 2 (MRP2) inhibitors. Eur. J. Pharm. Sci., 2009. **36**(4-5): p. 451-7.

307. Pedersen, J.M., Matsson, P., Bergstrom, C.A., Norinder, U., Hoogstraate, J., and Artursson, P., Prediction and identification of drug interactions with the human ATP-binding cassette transporter multidrug-resistance associated protein 2 (MRP2; ABCC2). J. Med. Chem., 2008. **51**(11): p. 3275-87.

308. Matsson, P., Englund, G., Ahlin, G., Bergstrom, C.A., Norinder, U., and Artursson, P., A global drug inhibition pattern for the human ATP-binding cassette transporter breast cancer resistance protein (ABCG2). J Pharmacol Exp Ther, 2007. **323**(1): p. 19-30.

309. Willett, P., Chemical Similarity Searching. J. Chem. Inf. Comput. Sci., 1998. **38**: p. 983-996.

310. Durant, J., Leland, B., Henry, D., and Nourse, J., Reoptimization of MDL keys for use in drug discovery. J. Chem. Inf. Comput. Sci., 2002. **42**: p. 1273-1280.

311. McGregor, M. and Pallai, P., Clustering of large databases of compounds: using the MDL “keys” as structural descriptors. J. Chem. Inf. Comput. Sci., 1997. **37**: p. 443-448.

312. Barnard, J., Clustering of Chemical Structures on the Basis of Two-Dimensional Similarity Measures. J. Chem. Inf. Comput. Sci., 1992. **32**: p. 644-649.

313. Willett, P., Similarity-based virtual screening using 2D fingerprints. Drug Discov Today, 2006. **11**(23-24): p. 1046-53.

314. Butina, D., Unsupervised data base clustering based on daylight's fingerprint and Tanimoto similarity: A fast and automated way to cluster small and large data sets. J. Chem. Inf. Comput. Sci., 1999. **39**: p. 747-750.

315. Glen, R. and Adams, S., Similarity Metrics and Descriptor Spaces – Which Combinations to Choose? QSAR Comb. Sci., 2006. **25**(12): p. 1133-1142.

316. Vapnik, V.N., The nature of statistical learning theory. 2000: springer.

317. Michielan, L. and Moro, S., Pharmaceutical Perspectives of Nonlinear QSAR Strategies. J. Chem. Inf. Comput. Sci., 2010. **50**: p. 961-978.

318. Shen, Q., Jiang, J., Tao, J., Shen, G., and Yu, R., Modified Ant Colony Optimization Algorithm for Variable Selection in QSAR

Modeling: QSAR Studies of Cyclooxygenase Inhibitors. J. Chem. Inf. Comput. Sci., 2005. **45**: p. 1024-1029.

319. Gunturi, S.B., Narayanan, R., and Khandelwal, A., In silico ADME modelling 2: computational models to predict human serum albumin binding affinity using ant colony systems. Bioorg Med Chem, 2006. **14**(12): p. 4118-29.

320. Goodarzi, M., Freitas, M.P., and Jensen, R., Feature selection and linear/nonlinear regression methods for the accurate prediction of glycogen synthase kinase-3beta inhibitory activities. J Chem Inf Model, 2009. **49**(4): p. 824-32.

321. Korb, O., Stutzle, T., and Exner, T.E., Empirical scoring functions for advanced protein-ligand docking with PLANTS. J Chem Inf Model, 2009. **49**(1): p. 84-96.

322. Izrailev, S. and Agrafiotis, D., A novel method for building regression tree models for QSAR based on artificial ant colony systems. J Chem Inf Comput Sci, 2001. **41**(1): p. 176-80.

323. Youden, W., Index For Rating Diagnostic Tests. Cancer, 1950. **1**: p. 32-35.

324. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. in Int. Joint Conf. Artif. Intell. 1995. Montreal, Quebec, Canada.

325. Knuth, D., Art of Computer Programming. 2011, Boston, USA: Addison-Wesley Professional.

326. Plouffe, D., Brinker, A., McNamara, C., Henson, K., Kato, N., Kuhen, K., Nagle, A., Adrian, F., Matzen, J.T., Anderson, P., Nam, T.G., Gray, N.S., Chatterjee, A., Janes, J., Yan, S.F., Trager, R., Caldwell, J.S., Schultz, P.G., Zhou, Y., and Winzeler, E.A., In silico activity profiling reveals the mechanism of action of antimalarials discovered in a high-throughput screen. Proc. Natl. Acad. Sci. USA, 2008. **105**(26): p. 9059-64.

327. Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E., and Willighagen, E., The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. J. Chem. Inf. Comput. Sci., 2003. **43**: p. 493-500.

328. Svetnik, V., Liaw, A., Tong, C., Culberson, J., Sheridan, R., and Feuston, B., Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. J. Chem. Inf. Comput. Sci., 2003. **43**: p. 1947-1958.

329. Cheng, T., Li, Q., Wang, Y., and Bryant, S.H., Binary classification of aqueous solubility using support vector machines with reduction and recombination feature selection. J Chem Inf Model, 2011. **51**(2): p. 229-36.

330. Todeschini, R. and Consonni, V., eds. Handbook of Molecular Descriptors. Methods and Principles in Medicinal Chemistry, ed. R. Mannhold, Kubinyi, H., and Timmerman, H. 2000, Wiley-VCH: Weinheim, New York.

331. Ramjee, M., Flinn, N., Pemberton, T., Quibell, M., Wang, Y., and Watts, J., Substrate mapping and inhibitor profiling of falcipain-2, falcipain-3 and berghepain-2: implications for peptidase anti-malarial drug discovery. Biochem. J., 2006. **399**: p. 47–57.

7.2 Curriculum vitae

Personal information:

Full name:	Claudia Suenderhauf
Date and place of birth:	3th of March 1980, Zurich, Switzerland
Nationality:	Swiss
Home address:	Streitgasse 2 4102 Binningen Telephone: 079 289 58 20 E-mail: claudia.suenderhauf@unibas.ch
Institutional address:	Pharmaceutical Technology Department of Pharmaceutical Sciences, University of Basel, Klingelbergstrasse 50 CH-4056 Basel, SWITZERLAND

Education:

1986 - 1994:	Primary School Buchs (SG), Switzerland
1994 - 1999:	Secondary School, Type B Sargans (SG) Switzerland
1999 - 2005:	Study of Medicine University of Basel Federal Swiss diploma in Human Medicine
2005 - 2006:	Assistant Physician in Clinical Trial Unit Swiss Pharma Contract, Allschwil, Switzerland
2006 - 2007:	Assistant Physician, Internal Medicine REHAB AG, Basel
2007:	Grant/Scholarship and inclusion in the national MD/PhD Program of the Swiss National Foundation Grant No. 323530-119218

2007 - 2008

MD/PhD Student
Developmental and Molecular Immunology
Zentrum für Biomedizin
University of Basel, Mattenstrasse 28
CH-4058 Basel, SWITZERLAND

2008 - now:

MD/PhD Student
Pharmaceutical Technology
Department of Pharmaceutical Sciences,
University of Basel, Klingelbergstrasse 50
CH-4056 Basel, SWITZERLAND

2010:

Medical thesis:
Cytokine signaling in the human brain capillary
endothelial cell line hCMEC/D3

2011:

PhD thesis:
In Silico Prediction of Drug Transport
Across Physiological Barriers

Publications:

Fasler-Kan E, **Suenderhauf C**, Barteneva N, Poller B, Gygax D, Huwyler J. *Cytokine signaling in the human brain capillary endothelial cell line hCMEC/D3*. Brain Res. 2010 Oct 1;1354:15-22.

Suenderhauf C, Hammann F, Maunz A, Helma C, Huwyler J. *Combinatorial QSAR modeling of human intestinal absorption*. Mol Pharm. 2011 Feb 7;8(1):213-24.

Hammann F, **Suenderhauf C**, Huwyler J. *A Binary Ant Colony Optimization Classifier for Molecular Activities. J Chem Inf Model. 2011 Sept 14; [Epub ahead of print].*

During my studies, I attended lectures by:

Arber S., Arvinte T., Beglinger C., Beier K., Brenner H.R., Drewe J., Eberle A., Engel J., Erb P., Ernst B., Fasler E., Finke D., Folkers G., Fricker G., Funk C,. Gehring J., Guentert T., Hauri H.-P., Hauser P., Hersberger K., Holländer G.A., Huwyler J., Imanidis G., Itin P., Jenal U., Ketman K., Körner C., Krähenbühl S., Leuenberger H., Melchers F., Otten U., Reichert H., Rolink A., Seelig A., Seelig J., Séquin U., Spiess M., Spornitz U., Vedani A., Vetter T. Vorobojev I.

Attended Meetings:

5th Scientific Meeting of the Swiss national MD-PhD Program, March 2008
Active participation, Poster presentation: **Claudia Suenderhauf**, Antonius Rolink, *The role of BAFF in human B cell development, homeostasis and disease.*

6th Scientific Meeting of the Swiss national MD-PhD Program, March 2010, Switzerland.
Active participation, Talk: **Claudia Suenderhauf**, Jörg Huwyler, *Introduction to data mining: Modeling Human oral Absorption*.

Annual Research Meeting of the Department of Pharmaceutical Sciences, Basel, Februrary 2011, Switzerland.
Active participation, Poster presentation: **Claudia Suenderhauf**, Felix Hammann, Andreas Maunz, Christoph Helma, and Jörg Huwyler. *Combinatorial QSAR Modeling of Human Intestinal Absorption.*

13. Herrenalber Transporter-Tage, Bad Herrenalb, May 2011, Germany

7.3 Supporting information Project 5.1

%Abs	Ordinal Class	SMILES
2	FALSE	CC1OC(OC2C(O)C(O)C(OC3C(O)C(O)C(O)OC3CO)OC2CO)C(O)C(O)C1NC4C=C(CO)C(O)C(O)C4O
50	UNKNOWN	CCCC(=O)Nc1ccc(OCC(O)CNC(C)C)c(c1)C(=O)C
90	TRUE	CC(=O)CC(c1ccc(cc1)[N+](=O)[O-])c2c(O)c3ccccc3oc2=O
95	TRUE	CC(=O)Nc1ccc(O)cc1
99	TRUE	CC(=O)Nc1nnc(s1)S(=O)(=O)N
90	TRUE	CC(=O)NC(CS)C(=O)O
90	TRUE	Cc1cnc(c[n+]1[O-])C(=O)O
90	TRUE	COc1cc(C)c(C=CC(=CC=CC(=CC(=O)O)C)C)c(C)c1C
17	FALSE	Nc1nc(O)c2ncn(COCCO)c2n1
85	TRUE	CC(C)(C)NCC(O)c1ccc(O)c(CO)c1
0	FALSE	OCC=C1C[N+]2(CC=C)CCC3(C2CC1C4=CN5C6C(=CN7C43)C8CC9C6(CC[N+]9(CC=C)CC8=CCO)c%10ccccc5%10)c%11ccccc7%11
1	FALSE	NCCCC(O)(P(=O)(O)O)P(=O)(O)O
100	TRUE	CC(C)CCCC(C)C1CCC2C(=CC=C3CC(O)CC(O)C3=C)CCCC12C
80	TRUE	O=c1[nH]cnc2[nH]ncc12
90	TRUE	Fc1ccc(cc1)C(N2CCN(CC2)c3nc(NCC=C)nc(NCC=C)n3)c4ccc(F)cc4
90	TRUE	NC1(CC2CC3CC(C2)C1)C3
10	FALSE	CC1(C)SC2C(N=CN3CCCCCC3)C(=O)N2C1C(=O)O
50	UNKNOWN	NC(=N)NC(=O)c1nc(Cl)c(N)nc1N
100	TRUE	CCC1(CCC(=O)NC1=O)c2ccc(N)cc2
50	UNKNOWN	CCCCc1oc2ccccc2c1C(=O)c3cc(I)c(OCCN(CC)CC)c(I)c3
95	TRUE	CN(C)CCC=C1c2ccccc2CCc3ccccc13
70	UNKNOWN	CCOC(=O)C1=C(COCCN)NC(=C(C1c2ccccc2Cl)C(=O)OC)C
95	TRUE	Clc1ccc2Oc3ccccc3N=C(N4CCNCC4)c2c1
90	TRUE	CC1(C)SC2C(NC(=O)C(N)c3ccc(O)cc3)C(=O)N2C1C(=O)O
35	UNKNOWN	CC1(C)SC2C(NC(=O)C(N)c3ccccc3)C(=O)N2C1C(=O)O
100	TRUE	CC(C)(C#N)c1cc(Cn2cncn2)cc(c1)C(C)(C)C#N
80	TRUE	CC(=O)Oc1ccccc1C(=O)O
90	TRUE	COc1ccc(CCN2CCC(CC2)Nc3nc4ccccc4n3Cc5ccc(F)cc5)cc1
44	UNKNOWN	CC(C)NCC(O)COc1ccc(CC(=O)N)cc1
95	TRUE	CN1C2CCC1CC(C2)OC(=O)C(CO)c3ccccc3
23	FALSE	CC(=O)OCC1OC(S)C(OC(=O)C)C(OC(=O)C)C1OC(=O)C
90	TRUE	CN1CCC(=C2c3ccccc3CCc4cccnc24)CC1
87	TRUE	Cn1cnc([N+](=O)[O-])c1Sc2[nH]cnc3ncnc23
37	UNKNOWN	CCC1OC(=O)C(C)C(OC2CC(C)(OC)C(O)C(C)O2)C(C)C(OC3OC(C)CC(C3O)N(C)C)C(C)(O)CC(C)CN(C)C(C)C(O)C1(C)O
0	FALSE	CC1(C)SC2C(NC(=O)C(NC(=O)N3CCNC3=O)c4ccccc4)C(=O)N2C1C(=O)O
1	FALSE	CC1C(NC(=O)C(=NOC(C)(C)C(=O)O)c2csc(N)n2)C(=O)N1S(=O)(=O)O
95	TRUE	NCC(CC(=O)O)c1ccc(Cl)cc1
37	UNKNOWN	CCOC(=O)C(CCc1ccccc1)NC2CCc3ccccc3N(CC(=O)O)C2=O
100	TRUE	NS(=O)(=O)c1cc2c(NC(Cc3ccccc3)NS2(=O)=O)cc1C(F)(F)F
100	TRUE	CC(=O)Nc1ccc(OC(=O)c2ccccc2OC(=O)C)cc1
70	UNKNOWN	NC(CO)C(=O)NNCc1ccc(O)c(O)c1O
100	TRUE	CNCCc1ccccn1
85	TRUE	CC(C)NCC(O)COc1ccc(CCOCC2CC2)cc1
100	TRUE	CC(C)(Oc1ccc(CCNC(=O)c2ccc(Cl)cc2)cc1)C(=O)O

95	TRUE	CNCCCCOc1ccccc1Cc2ccccc2
100	TRUE	OC(CCN1CCCCC1)(C2CC3CC2C=C3)c4ccccc4
90	TRUE	CC(C)NCC(O)COc1ccc(COCCOC(C)C)cc1
35	UNKNOWN	CC(C)CC1N2C(=O)C(NC(=O)C3CN(C)C4Cc5c(Br)[nH]c6cccc(C4=C3)c56)(OC2(O)C7CCCN7C1=O)C(C)C
50	UNKNOWN	CN(C)CCC(c1ccc(Br)cc1)c2ccccn2
100	TRUE	CCCC1OC2CC3C4CCC5=CC(=O)C=CC5(C)C4C(O)CC3(C)C2(O1)C(=O)CO
80	TRUE	CCCCNc1cc(cc(c1Oc2ccccc2)S(=O)(=O)N)C(=O)O
95	TRUE	CC(NC(C)(C)C)C(=O)c1cccc(Cl)c1
100	TRUE	O=C1CC2(CCCC2)CC(=O)N1CCCCN3CCN(CC3)c4ncccn4
95	TRUE	Cn1cnc2n(C)c(=O)n(C)c(=O)c12
100	TRUE	CC(CCCC(C)(C)O)C1CCC2C(=CC=C3CC(O)CC(O)C3=C)CCCC12C
72	UNKNOWN	CC(CS)C(=O)N1CCCC1C(=O)O
80	TRUE	NC(=O)N1c2ccccc2C=Cc3ccccc13
90	TRUE	CC1(C)C(CCC2(C)C1CCC3(C)C2C(=O)C=C4C5CC(C)(CCC5(C)CCC43C)C(=O)O)OC(=O)CCC(=O)O
60	UNKNOWN	CC(Cc1ccc(O)c(O)c1)(NN)C(=O)O
95	TRUE	CCOC(=O)n1ccn(C)c1=S
95	TRUE	CC(C)(C)NCC(O)COc1cccc2NC(=O)CCc12
80	TRUE	COc1ccccc1OCCNCC(O)COc2cccc3[nH]c4ccccc4c23
60	UNKNOWN	Nc1nc(cs1)C(=NOCC(=O)O)C(=O)NC2C3SCC(=C(N3C2=O)C(=O)O)C=C
0	FALSE	CON=C(C(=O)NC1C2SCC(=C(N2C1=O)C(=O)O)CSc3nc(C)c(CC(=O)O)s3)c4csc(N)n4
0	FALSE	COC1(NC(=O)Cc2cccs2)C3SCC(=C(N3C1=O)C(=O)O)COC(=O)N
50	UNKNOWN	COCC1=C(N2C(SC1)C(NC(=O)C(=NOC)c3csc(N)n3)C2=O)C(=O)OC(C)OC(=O)OC(C)C
94	TRUE	CC=CC1=C(N2C(SC1)C(NC(=O)C(N)c3ccc(O)cc3)C2=O)C(=O)O
0	FALSE	CON=C(C(=O)NC1C2SCC=C(N2C1=O)C(=O)O)c3csc(N)n3
55	UNKNOWN	CCN(CC)C(=O)Nc1ccc(OCC(O)CNC(C)(C)C)c(c1)C(=O)C
90	TRUE	CC1=C(N2C(SC1)C(NC(=O)C(N)c3ccccc3)C2=O)C(=O)O
100	TRUE	OC(=O)COCCN1CCN(CC1)C(c2ccccc2)c3ccc(Cl)cc3
100	TRUE	CC(CCC(=O)O)C1CCC2C1(CCC3C2C(CC4C3(CCC(C4)O)C)O)C
95	TRUE	C[N+](C)(C)CC(=O)[O-]
95	TRUE	OC(O)C(Cl)(Cl)Cl
85	TRUE	OCC(NC(=O)C(Cl)Cl)C(O)c1ccc(cc1)[N+](=O)[O-]
95	TRUE	CNC1=Nc2ccc(Cl)cc2C(=[N+]([O-])C1)c3ccccc3
100	TRUE	CCN(CC)CCCC(C)Nc1ccnc2cc(Cl)ccc12
20	FALSE	NS(=O)(=O)c1cc2c(N=CNS2(=O)=O)cc1Cl
80	TRUE	CN(C)CCC(c1ccc(Cl)cc1)c2ccccn2
96	TRUE	CN(C)CCCN1c2ccccc2Sc3ccc(Cl)cc13
100	TRUE	CCCNC(=O)NS(=O)(=O)c1ccc(Cl)cc1
60	UNKNOWN	CN(C)C1C2CC3C(=C(O)C2(O)C(=O)C(=C1O)C(=O)N)C(=O)c4c(O)ccc(Cl)c4C3(C)O
65	UNKNOWN	NS(=O)(=O)c1cc(ccc1Cl)C2(O)NC(=O)c3ccccc32
60	UNKNOWN	CCOC(=O)C(CCc1ccccc1)NC2CCCN3CCCC(N3C2=O)C(=O)O
90	TRUE	CNC(=NC#N)NCCSCc1nc[nH]c1C
95	TRUE	CCn1nc(C(=O)O)c(=O)c2cc3OCOc3cc12
99	TRUE	CC(C)(Oc1ccc(cc1)C2CC2(Cl)Cl)C(=O)O
77	UNKNOWN	OC(=O)c1cn(C2CC2)c3cc(N4CCNCC4)c(F)cc3c1=O
95	TRUE	COC1CN(CCCOc2ccc(F)cc2)CCC1NC(=O)c3cc(Cl)c(N)cc3OC
100	TRUE	CN(C)CCCC1(OCc2cc(C#N)ccc21)c3ccc(F)cc3
100	TRUE	CCC1OC(=O)C(C)C(OC2CC(C)(OC)C(O)C(C)O2)C(C)C(OC3OC(C)CC(C3O)N(C)C)C(C)(CC(C)C(=O)C(C)C(O)C1(C)O)OC
75	UNKNOWN	OCC=C1OC2CC(=O)N2C1C(=O)O
95	TRUE	CN1C(=O)CC(=O)N(c2ccccc2)c3cc(Cl)ccc13

95	TRUE	CCOC(=O)C(C)(C)Oc1ccc(Cl)cc1
95	TRUE	Cc1ncsc1CCCl
95	TRUE	CN(C)CCCN1c2ccccc2CCc3ccc(Cl)cc13
80	TRUE	[O-][N+](=O)c1ccc2NC(=O)CN=C(c3ccccc3Cl)c2c1
100	TRUE	Clc1cccc(Cl)c1N=C2NCCN2
100	TRUE	CC1CCCC(C)N1NC(=O)c2ccc(Cl)c(c2)S(=O)(=O)N
93	TRUE	CN1CCN(CC1)C2=Nc3cc(Cl)ccc3Nc4ccccc24
90	TRUE	Coc1ccc2CC3C4C=CC(O)C5Oc1c2C54CCN3C
100	TRUE	COc1cc2CCC(NC(=O)C)c3cc(=O)c(OC)ccc3-c2c(OC)c1OC
95	TRUE	CC(=O)OCC(=O)C1(O)CCC2C3CCC4=CC(=O)CCC4(C)C3C(=O)CC21C
1	FALSE	OC(COc1cccc2oc(cc(=O)c12)C(=O)O)COc3cccc4oc(cc(=O)c34)C(=O)O
75	UNKNOWN	ClCCN(CCCl)P1(=O)NCCCO1
100	TRUE	NC1CONC1=O
40	UNKNOWN	CCC1NC(=O)C(C(O)C(C)CC=CC)N(C)C(=O)C(C(C)C)N(C)C(=O)C(CC(C)C)N(C)C(=O)C(CC(C)C)N(C)C(=O)C(C)NC(=O)C(C)NC(=O)C(CC(C)C)N(C)C(=O)C(NC(=O)C(CC(C)C)N(C)C(=O)CN(C)C1=O)C(C)C
95	TRUE	CC(=O)OC1(CCC2C3C=C(Cl)C4=CC(=O)C5CC5C4(C)C3CCC21C)C(=O)C
80	TRUE	[O-][N+](=O)c1ccc(cc1)c2ccc(C=NN3CC(=O)NC3=O)o2
90	TRUE	Nc1ccc(cc1)S(=O)(=O)c2ccc(N)cc2
0	FALSE	Coc1cccc2C(=O)c3c(O)c4CC(O)(CC(OC5CC(N)C(O)C(C)O5)c4c(O)c3C(=O)c12)C(=O)C
90	TRUE	CNCCCN1c2ccccc2CCc3ccccc13
70	UNKNOWN	CCC12CC(=C)C3C(CCC4=CCCCC34)C2CCC1(O)C#C
83	TRUE	CCNC(C)Cc1cccc(c1)C(F)(F)F
100	TRUE	CN1CCC23C4Oc5c3c(CC1C2C=CC4OC(=O)C)ccc5OC(=O)C
100	TRUE	CN1C(=O)CN=C(c2ccccc2)c3cc(Cl)ccc13
90	TRUE	CC1=Nc2ccc(Cl)cc2S(=O)(=O)N1
90	TRUE	OC(=O)Cc1ccccc1Nc2c(Cl)cccc2Cl
30	FALSE	OCC1CCC(O1)n2cnc3c(O)ncnc23
90	TRUE	CCN(CC)C(=O)N1CCN(CC1)C
90	TRUE	CCC(=C(CC)c1ccc(O)cc1)c2ccc(O)cc2
90	TRUE	OC(=O)c1cc(ccc1O)c2ccc(F)cc2F
90	TRUE	CC1OC(CC(O)C1O)OC2C(C)OC(CC2O)OC3C(C)OC(CC3O)OC4CCC5(C)C(CCC6C5CCC7(C)C(CCC67O)C8=CC(=O)OC8)C4
80	TRUE	CC1OC(CC(O)C1O)OC2C(C)OC(CC2O)OC3C(C)OC(CC3O)OC4CCC5(C)C(CCC6C5CC(O)C7(C)C(CCC67O)C8=CC(=O)OC8)C4
97	TRUE	COc1ccc2CC3C4CCC(O)C5Oc1c2C54CCN3C
90	TRUE	CN(C(=O)C(Cl)Cl)c1ccc(OC(=O)c2ccco2)cc1
90	TRUE	Coc1ccc(cc1)C2Sc3ccccc3N(CCN(C)C)C(=O)C2OC(=O)C
90	TRUE	CN(C)CCOC(c1ccccc1)c2ccccc2
3	FALSE	OP(=O)([O-])C(Cl)(Cl)P(=O)(O)[O-]
95	TRUE	CC(C)N(CCC(C(=O)N)(c1ccccc1)c2ccccn2)C(C)C
80	TRUE	CCN(CC)C(=S)SSC(=S)N(CC)CC
78	UNKNOWN	O=C(OC1CC2CC3CC(C1)N2CC3=O)c4c[nH]c5ccccc45
93	TRUE	Clc1ccc2n(C3CCN(CCCn4c(=O)[nH]c5ccccc45)CC3)c(=O)[nH]c2c1
0	FALSE	NCCc1ccc(O)c(O)c1
95	TRUE	CN(C)CCC=C1c2ccccc2CSc3ccccc13
65	UNKNOWN	Coc1cc2nc(nc(N)c2cc1OC)N3CCN(CC3)C(=O)C4COc5ccccc5O4
100	TRUE	CN(C)CCC=C1c2ccccc2COc3ccccc13
93	TRUE	CC1C2C(O)C3C(N(C)C)C(=C(C(=O)N)C(=O)C3(O)C(=C2C(=O)c4c(O)cccc14)O)O
55	UNKNOWN	NCCCC(N)(C(F)F)C(=O)O
60	UNKNOWN	CCOC(=O)C(CCc1ccccc1)NC(C)C(=O)N2CCCC2C(=O)O
80	TRUE	CSc1ccc(cc1)C(=O)c2[nH]c(=O)[nH]c2C

100	TRUE	CNC(C)C(O)c1ccccc1
0	FALSE	CC1C(C(CC(O1)OC2CC(CC3=C(C4=C(C(=C23)O)C(=O)C5=C(C4=O)C=CC=C5OC)O)(C(=O)CO)O)N)O
90	TRUE	CCC(=C)C(=O)c1ccc(OCC(=O)O)c(Cl)c1Cl
80	TRUE	CCC(CO)NCCNC(CC)CO
90	TRUE	CCc1cc(ccn1)C(=S)N
100	TRUE	CCC1(C)CC(=O)NC1=O
5	FALSE	CC(O)(P(=O)(O)O)P(=O)(O)O
73	UNKNOWN	CCc1cccc2c3CCOC(CC)(CC(=O)O)c3[nH]c12
40	UNKNOWN	CCOC(=O)C=C(C)C=CC=C(C)C=Cc1c(C)cc(OC)c(C)c1C
73	UNKNOWN	CC(=O)OCC(CCn1cnc2cnc(N)nc12)COC(=O)C
40	UNKNOWN	NC(=Nc1nc(CSCCC(=NS(=O)(=O)N)N)cs1)N
90	TRUE	NC(=O)OCC(COC(=O)N)c1ccccc1
100	TRUE	CCOC(=O)C1=C(C)NC(=C(C1c2cccc(Cl)c2Cl)C(=O)OC)C
80	TRUE	OC(=O)CCC(=O)c1ccc(cc1)c2ccccc2
95	TRUE	CCNC(C)Cc1cccc(c1)C(F)(F)F
75	UNKNOWN	CC(C)OC(=O)C(C)(C)Oc1ccc(cc1)C(=O)c2ccc(Cl)cc2
85	TRUE	CC(C(=O)O)c1cccc(Oc2ccccc2)c1
100	TRUE	CC(C)(C)NC(=O)C1CCC2C3CCC4NC(=O)C=CC4(C)C3CCC12C
95	TRUE	FC(F)(F)COc1ccc(OCC(F)(F)F)c(c1)C(=O)NCC2CCCCN2
80	TRUE	Cc1onc(c1C(=O)NC2C3SC(C)(C)C(N3C2=O)C(=O)O)c4c(F)cccc4Cl
90	TRUE	OC(Cn1cncn1)(Cn2cncn2)c3ccc(F)cc3F
100	TRUE	Nc1nc(=O)[nH]cc1F
75	UNKNOWN	Nc1nc(F)nc2n(cnc12)C3OC(COP(=O)(O)O)C(O)C3O
95	TRUE	CC(=O)OCC(=O)C1(O)CCC2C3CCC4=CC(=O)CCC4(C)C3(F)C(O)CC21C
95	TRUE	CCOC(=O)c1ncn-2c1CN(C)C(=O)c3cc(F)ccc32
95	TRUE	Fc1ccc(cc1)C(N2CCN(CC=Cc3ccccc3)CC2)c4ccc(F)cc4
80	TRUE	CC1(C)OC2CC3C4CC(F)C5=CC(=O)C=CC5(C)C4C(O)CC3(C)C2(O1)C(=O)CO
28	FALSE	Fc1c[nH]c(=O)[nH]c1=O
95	TRUE	CNCCC(Oc1ccc(cc1)C(F)(F)F)c2ccccc2
100	TRUE	OCCN1CCN(CCC=C2c3ccccc3Sc4ccc(cc24)C(F)(F)F)CC1
100	TRUE	CCC(CC)CCN1C(=O)CN=C(C2CCCCC2F)c3cc(Cl)ccc13
95	TRUE	CC(C(=O)O)c1ccc(c(F)c1)c2ccccc2
90	TRUE	CC(C)C(=O)Nc1ccc([N+](=O)[O-])c(c1)C(F)(F)F
20	FALSE	CCC(=O)OC1(C(C)CC2C3CC(F)C4=CC(=O)C=CC4(C)C3(F)C(O)CC21C)C(=O)SCF
95	TRUE	CC(C)n1c(C=CC(O)CC(O)CC(=O)O)c(c2ccc(F)cc2)c3ccccc13
90	TRUE	COCCCCC(=NOCCN)c1ccc(cc1)C(F)(F)F
75	UNKNOWN	Nc1nc(=O)c2nc(CNc3ccc(cc3)C(=O)NC(CCC(=O)O)C(=O)O)cnc2[nH]1
65	UNKNOWN	Coc1ccc(CC(C)NCC(O)c2ccc(O)c(NC=O)c2)cc1
20	FALSE	OC(=O)P(=O)(O)O
50	UNKNOWN	CC1OC1P(=O)(O)O
34	UNKNOWN	CCC(=O)OC(OP(=O)(CCCCc1ccccc1)CC(=O)N2CC(CC2C(=O)O)C3CCCCC3)C(C)C
65	UNKNOWN	NS(=O)(=O)c1cc(C(=O)O)c(NCc2ccco2)cc1Cl
100	TRUE	CC1C(O)CCC2(C)C1CCC3(C)C2C(O)CC4C(=C(CCC=C(C)C)C(=O)O)C(CC43C)OC(=O)C
60	UNKNOWN	NCC1(CC(=O)O)CCCCC1
90	TRUE	Coc1ccc(CCN(C)CCCC(C#N)(C(C)C)c2cc(OC)c(OC)c(OC)c2)cc1OC
5	FALSE	Nc1nc(O)c2ncn(COC(CO)CO)c2n1
100	TRUE	Cc1ccc(C)c(OCCCC(C)(C)C(=O)O)c1
100	TRUE	CCC12CCC3C(CCC4=CC(=O)CCC34)C2C=CC1(O)C#C
60	UNKNOWN	CCC12C=CC3=C4CCC(=O)C=C4CCC3C2CCC1(O)C#C
98	TRUE	Cc1ccc(cc1)S(=O)(=O)NC(=O)NC2C(O)C3(C)CCC2C3(C)C

100	TRUE	Cc1cnc(cn1)C(=O)NCCc2ccc(cc2)S(=O)(=O)NC(=O)NC3CCCCC3
95	TRUE	COc1ccc2c(c1)C(=O)N(CCc3ccc(cc3)S(=O)(=O)NC(=O)NC4CCCCC4)C(=O)C2(C)C
95	TRUE	COc1ccc(Cl)cc1C(=O)NCCc2ccc(cc2)S(=O)(=O)NC(=O)NC3CCCCC3
17.5	FALSE	C[N+]1(C)CCC(C1)OC(=O)C(O)(C2CCCC2)c3ccccc3
95	TRUE	COCCOc1cnc(NS(=O)(=O)c2ccccc2)nc1
100	TRUE	CN1C2CCCC1CC(C2)NC(=O)c3nn(C)c4ccccc34
100	TRUE	OC1(CCN(CCCC(=O)c2ccc(F)cc2)CC1)c3ccc(Cl)cc3
100	TRUE	Nnc1nncc2ccccc12
70	UNKNOWN	NS(=O)(=O)c1cc2c(NCNS2(=O)=O)cc1Cl
90	TRUE	CCN(CCO)CCCC(C)Nc1ccnc2cc(Cl)ccc12
90	TRUE	CCCCCC(=O)OC1(CCC2C3CCC4=CC(=O)CCC4(C)C3CCC21C)C(=O)C
95	TRUE	CC(C)Cc1ccc(cc1)C(C)C(=O)O
82	TRUE	CCCCCCCN(CC)CCCC(O)c1ccc(NS(=O)(=O)C)cc1
30	FALSE	CC1OC(CC(N)C1O)OC2CC(O)(Cc3c(O)c4C(=O)c5ccccc5C(=O)c4c(O)c23)C(=O)C
0	FALSE	OCC1OC(CC1O)n2cc(I)c(=O)[nH]c2=O
16	FALSE	CC#CCC(C)C(O)C=CC1C(O)CC2CC(=CCCCC(=O)O)CC12
5	FALSE	CC(O)C1C2CC(=C(N2C1=O)C(=O)O)SCCNC=N
90	TRUE	CN(C)CCCN1c2ccccc2CCc3ccccc13
19	FALSE	CC(C)(C)NC(=O)C1CN(Cc2cccnc2)CCN1CC(O)CC(Cc3ccccc3)C(=O)NC4C(O)Cc5ccccc45
100	TRUE	COc1ccc2n(C(=O)c3ccc(Cl)cc3)c(C)c(CC(=O)O)c2c1
95	TRUE	NNC(=O)c1ccncc1
85	TRUE	CC(C)NCC(O)c1ccc(O)c(O)c1
90	TRUE	CC(=CC=CC(=CC(=O)O)C)C=CC1=C(C)CCCC1(C)C
90	TRUE	COC(=O)C1=C(C)NC(=C(C1c2cccc3nonc23)C(=O)OC(C)C)C
85	TRUE	CCC(C)n1ncn(c2ccc(cc2)N3CCN(CC3)c4ccc(OCC5COC(Cn6cncn6)(O5)c7ccc(Cl)cc7Cl)cc4)c1=O
60	UNKNOWN	COC1CC(OC2C(C)OC(CC2OC)OC3C(C)C=CC=C4COC5C(O)C(=CC(C(=O)OC(C)CC6(CCC(C)CO6)OC(C)CC=C3C)C54O)C)OC(C)C1O
100	TRUE	Fc1ccc(cc1)C(=O)C2CCN(CCn3c(=O)[nH]c4ccccc4c3=O)CC2
100	TRUE	CN1C(=O)CN2C(=O)C=C(C)OC2(c3ccccc3)c4cc(Cl)ccc14
90	TRUE	CC(C(=O)O)c1cccc(c1)C(=O)c2ccccc2
95	TRUE	OC(=O)C1CCn2c(ccc12)C(=O)c3ccccc3
90	TRUE	CN1CCC(=C2c3ccsc3C(=O)Cc4ccccc24)CC1
95	TRUE	CCOC(=O)C1=C(C)NC(=C(C1c2ccccc2C=CC(=O)OC(C)(C)C)C(=O)OCC)C
2	FALSE	OCC1OC(O)(CO)C(O)C1OC2OC(CO)C(O)C(O)C2O
98	TRUE	Nc1nnc(c(N)n1)c2cccc(Cl)c2Cl
85	TRUE	Cc1c(CS(=O)c2nc3ccccc3[nH]2)nccc1OCC(F)(F)F
95	TRUE	C1CN2CC(N=C2S1)c3ccccc3
100	TRUE	CC(C)(C)NCC(O)COc1cccc2C(=O)CCCc12
95	TRUE	CCN(CC)CC(=O)Nc1c(C)cccc1C
25	FALSE	NCCCCC(NC(CCc1ccccc1)C(=O)O)C(=O)N2CCCC2C(=O)O
100	TRUE	CCN(CC)C(=O)NC1CN(C)C2Cc3c[nH]c4cccc(C2=C1)c34
75	UNKNOWN	OC(=O)C(=O)Nc1cc(C#N)cc(NC(=O)C(=O)O)c1Cl
98	TRUE	Ccn1cc(C(=O)O)c(=O)c2cc(F)c(N3CCNC(C)C3)c(F)c12
65	UNKNOWN	CN(C)C(=O)C(CCN1CCC(O)(CC1)c2ccc(Cl)cc2)(c3ccccc3)c4ccccc4
90	TRUE	CCOC(=O)N1CCC(=C2c3ccc(Cl)cc3CCc4cccnc24)CC1
90	TRUE	OC1N=C(c2ccccc2Cl)c3cc(Cl)ccc3NC1=O
66	UNKNOWN	CCCCc1nc(Cl)c(CO)n1Cc2ccc(cc2)c3ccccc3c4nnn[nH]4
95	TRUE	CNCCCC1(CCC2c3ccccc31)c4ccccc24
7.5	FALSE	COC(=O)Nc1nc2cc(ccc2[nH]1)C(=O)c3ccccc3
90	TRUE	CCN(CCCCOC(=O)c1ccc(OC)c(OC)c1)C(C)Cc2ccc(OC)cc2
90	TRUE	Cc1cccc(Nc2ccccc2C(=O)O)c1C

77.5	UNKNOWN	OC(C1CCCCN1)c2cc(nc3c(cccc23)C(F)(F)F)C(F)(F)F
100	TRUE	CC(=O)OC1(CCC2C3C=C(C)C4=CC(=O)CCC4(C)C3CCC21C)C(=O)C
100	TRUE	CCOC(=O)C1(CCN(C)CC1)c2ccccc2
100	TRUE	CCC1(CCCCN(C)C1)c2cccc(O)c2
0	FALSE	CC(O)C1C2C(C)C(=C(N2C1=O)C(=O)O)SC3CNC(C3)C(=O)N(C)C
90	TRUE	COc1ccc2C3CCC4(C)C(CCC4(O)C#C)C3CCc2c1
55	UNKNOWN	CN(C)C(=N)NC(=N)N
95	TRUE	Cn1cc[nH]c1=S
95	TRUE	CC(CO)NC(=O)C1CN(C)C2Cc3cn(C)c4cccc(C2=C1)c34
80	TRUE	COC(=O)C(C1CCCCN1)c2ccccc2
90	TRUE	CC1CC2C3CCC(O)(C(=O)COC(=O)CCC(=O)O)C3(C)CC(O)C2C4(C)C=CC(=O)C=C14
100	TRUE	CCC(CO)NC(=O)C1CN(C)C2Cc3cn(C)c4cccc(C2=C1)c34
90	TRUE	CCN(CC)CCNC(=O)c1cc(Cl)c(N)cc1OC
64	UNKNOWN	CC1Nc2cc(Cl)c(cc2C(=O)N1c3ccccc3C)S(=O)(=O)N
95	TRUE	COCCc1ccc(OCC(O)CNC(C)C)cc1
95	TRUE	Cc1ncc([N+](=O)[O-])n1CCO
0	FALSE	CC1(C)SC2C(NC(=O)C(NC(=O)N3CCN(C3=O)S(=O)(=O)C)c4ccccc4)C(=O)N2C1C(=O)O
70	UNKNOWN	CN1CCN2C(C1)c3ccccc3Cc4ccccc24
100	TRUE	COCC(=O)OC1(CCN(C)CCCc2nc3ccccc3[nH]2)CCc4cc(F)ccc4C1C(C)C
20	FALSE	Clc1ccc(COC(Cn2ccnc2)c3ccc(Cl)cc3Cl)c(Cl)c1
100	TRUE	Cc1ncc2CN=C(c3ccccc3F)c4cc(Cl)ccc4-n12
90.6	TRUE	CC#CC1(O)CCC2C3CCC4=CC(=O)CCC4=C3C(CC21C)c5ccc(cc5)N(C)C
82.5	TRUE	CC1=NC(=O)C(C=C1c2ccncc2)C#N
100	TRUE	CN(C)C1C2CC3Cc4c(ccc(O)c4C(=O)C3=C(O)C2(O)C(=O)C(=C1O)C(=O)N)N(C)C
100	TRUE	Nc1cc(nc(N)[n+]1[O-])N2CCCCC2
88	TRUE	CCCCC(C)(O)CC=CC1C(O)CC(=O)C1CCCCCCC(=O)OC
95	TRUE	Clc1ccc(cc1)C(=O)NCCN2CCOCC2
100	TRUE	CCOC(=O)N=c1c[n+]([n-]o1)N2CCOCC2
95	TRUE	CCOC(=O)Nc1ccc2Sc3ccccc3N(C(=O)CCN4CCOCC4)c2c1
37.5	UNKNOWN	CN1CCC23C4Oc5c3c(CC1C2C=CC4O)ccc5O
0	FALSE	COC1(NC(=O)C(C(=O)O)c2ccc(O)cc2)C3OCC(=C(N3C1=O)C(=O)O)CSc4nnnn4C
80	TRUE	COc1ccc2cc(CCC(=O)C)ccc2c1
30	FALSE	CC(C)(C)NCC(O)COc1cccc2CC(O)C(O)Cc12
92.5	TRUE	CCN(CC)CCOC(=O)C(CC1CCCO1)Cc2cccc3ccccc23
100	TRUE	OC1CCC2(O)C3Cc4ccc(O)c5OC1C2(CCN3CC6CCC6)c54
90	TRUE	CCn1cc(C(=O)O)c(=O)c2ccc(C)nc12
95	TRUE	Oc1ccc2CC3N(CC=C)CCC4(C5Oc1c24)C3(O)CCC5=O
100	TRUE	Oc1ccc2CC3N(CC4CC4)CCC5(C6Oc1c25)C3(O)CCC6=O
100	TRUE	Coc1ccc2cc(ccc2c1)C(C)C(=O)O
2.5	FALSE	CCCc1c2oc(cc(=O)c2cc3c(=O)cc(C(=O)O)n(CC)c13)C(=O)O
99	TRUE	CCc1nn(CCCN2CCN(CC2)c3cccc(Cl)c3)c(=O)n1CCOc4ccccc4
97.5	TRUE	CN1CCOC(c2ccccc2)c3ccccc3C1
5	FALSE	NCC1OC(OC2C(N)CC(N)C(O)C2O)C(N)C(O)C1O
1.5	FALSE	CN(C)C(=O)Oc1cccc(c1)[N+](C)(C)C
95	TRUE	COC(=O)C1=C(C)NC(=C(C1c2cccc(c2)[N+](=O)[O-])C(=O)OCCN(C)Cc3ccccc3)C
100	TRUE	[O-][N+](=O)OCCNC(=O)c1cccnc1
100	TRUE	CN1CCCC1c2cccnc2
90	TRUE	COC(=O)C1=C(C)NC(=C(C1c2ccccc2[N+](=O)[O-])C(=O)OC)C
53	UNKNOWN	COCCOC(=O)C1=C(C)NC(=C(C1c2cccc(c2)[N+](=O)[O-])C(=O)OC(C)C)C
100	TRUE	COC(=O)C1=C(C)NC(=C(C1c2ccccc2[N+](=O)[O-])C(=O)OCC(C)C)C

95	TRUE	[O-][N+](=O)c1ccc2NC(=O)CN=C(c3ccccc3)c2c1
95	TRUE	[O-][N+](=O)c1ccc(C=NN2CC(=O)NC2=O)o1
70	UNKNOWN	CNC(=C[N+](=O)[O-])NCCSCc1csc(CN(C)C)n1
100	TRUE	CCC12CCC3C(CCC4=CC(=NO)CCC34)C2CCC1(OC(=O)C)C#C
100	TRUE	CNCCC=C1c2ccccc2CCc3ccccc13
100	TRUE	CC1COc2c(N3CCN(C)CC3)c(F)cc4c(=O)c(cn1c24)C(=O)O
3	FALSE	OC(=O)c1cc(N=Nc2ccc(O)c(c2)C(=O)O)ccc1O
65	UNKNOWN	COc1ccc2[nH]c(nc2c1)S(=O)Cc3ncc(C)c(OC)c3C
59	UNKNOWN	Cc1nccn1CC2CCc3c(C2=O)c4ccccc4n3C
100	TRUE	CN(C)CCOC(c1ccccc1)c2ccccc2C
5	FALSE	CC1OC(OC2CC(O)C3(CO)C4C(O)CC5(C)C(CCC5(O)C4CCC3(O)C2)C6=CC(=O)OC6)C(O)C(O)C1O
98	TRUE	OC1N=C(c2ccccc2)c3cc(Cl)ccc3NC1=O
16	FALSE	CC[N+]1(C)C2CC(CC1C3OC32)OC(=O)C(CO)c4ccccc4
100	TRUE	CC(C)NCC(O)COc1ccccc1OCC=C
100	TRUE	CCN(CC)CC#CCOC(=O)C(O)(C1CCCCC1)c2ccccc2
60	UNKNOWN	CN(C)C1C2C(O)C3C(=C(O)C2(O)C(=O)C(=C1O)C(=O)N)C(=O)c4c(O)cccc4C3(C)O
0	FALSE	CC(=O)OC1C(CC2C3CCC4CC(OC(=O)C)C(CC4(C)C3CCC12C)[N+]5(C)CCCCC5)[N+]6(C)CCCCC6
77	UNKNOWN	COc1ccnc(CS(=O)c2nc3cc(OC(F)F)ccc3[nH]2)c1OC
90	TRUE	COc1ccc(Cc2nccc3cc(OC)c(OC)cc23)cc1OC
100	TRUE	Fc1ccc(cc1)C2CCNCC2COc3ccc4OCOc4c3
100	TRUE	CC(C)(C)NCC(C)(O)COc1ccccc1C2CCCC2
40	UNKNOWN	CC(C)(S)C(N)C(=O)O
30	FALSE	CC1(C)SC2C(NC(=O)Cc3ccccc3)C(=O)N2C1C(=O)O
0	FALSE	NC(=N)c1ccc(OCCCCCOc2ccc(cc2)C(=N)N)cc1
100	TRUE	CC1C2Cc3ccc(O)cc3C1(C)CCN2CC=C(C)C
100	TRUE	CCCC(C)C1(CC)C(=O)NC(=O)NC1=O
95	TRUE	CC(=O)CCCCn1c(=O)n(C)c2ncn(C)c2c1=O
60	UNKNOWN	CCCN1CC(CSC)CC2C1Cc3c[nH]c4cccc2c34
95	TRUE	CCCC(NC(C)C(=O)N1C(CC2CCCCC21)C(=O)O)C(=O)OCC
100	TRUE	OCCN1CCN(CCCN2c3ccccc3Sc4ccc(Cl)cc24)CC1
100	TRUE	O=C1C(C(=O)c2ccccc12)c3ccccc3
90	TRUE	CCC1(C(=O)NC(=O)NC1=O)c2ccccc2
12	FALSE	CCOC(=O)C1(CCN(CCC(O)c2ccccc2)CC1)c3ccccc3
25	FALSE	CC(COc1ccccc1)N(CCCl)Cc2ccccc2
95	TRUE	CCC(c1ccccc1)c2c(O)c3ccccc3oc2=O
95	TRUE	CCCCC1C(=O)N(N(C1=O)c2ccccc2)c3ccccc3
90	TRUE	CC(N)C(O)c1ccccc1
90	TRUE	O=C1NC(=O)C(N1)(c2ccccc2)c3ccccc3
70	UNKNOWN	Fc1ccc(cc1)C(CCCN2CCC(CC2)n3c(=O)[nH]c4ccccc34)c5ccc(F)cc5
90	TRUE	CC(C)NCC(O)COc1cccc2[nH]ccc12
25	FALSE	CN1CCN(CC(=O)N2c3ccccc3C(=O)Nc4cccnc24)CC1
86	TRUE	NS(=O)(=O)c1cc(cc(N2CCCC2)c1Oc3ccccc3)C(=O)O
100	TRUE	CN1C(=C(O)c2ccccc2S1(=O)=O)C(=O)Nc3ccccn3
80	TRUE	CN1CCC(=C2c3ccsc3CCc4ccccc24)CC1
100	TRUE	CN1C(CSCC(F)(F)F)Nc2cc(Cl)c(cc2S1(=O)=O)S(=O)(=O)N
30	FALSE	C[n+]1ccccc1C=NO
34	UNKNOWN	CCC(C)C(=O)OC1CC(O)C=C2C=CC(C)C(CCC(O)CC(O)CC(=O)O)C12
90	TRUE	O=C(C1CCCCC1)N2CC3N(CCc4ccccc34)C(=O)C2
57	UNKNOWN	COc1cc2nc(nc(N)c2cc1OC)N3CCN(CC3)C(=O)c4ccco4

75	UNKNOWN	CCC1(C(=O)NCNC1=O)c2ccccc2
100	TRUE	CCCN(CCC)S(=O)(=O)c1ccc(cc1)C(=O)O
82.5	TRUE	CCN(CC)CCNC(=O)c1ccc(N)cc1
95	TRUE	CNNCc1ccc(cc1)C(=O)NC(C)C
100	TRUE	OC(CCN1CCCC1)(C2CCCCC2)c3ccccc3
90	TRUE	CC(C)NC(=N)NC(=N)Nc1ccc(Cl)cc1
80	TRUE	CC(CN1c2ccccc2Sc3ccccc13)N(C)C
95	TRUE	CCCNCC(O)COc1ccccc1C(=O)CCc2ccccc2
10	FALSE	CC(C)[N+](C)(CCOC(=O)C1c2ccccc2Oc3ccccc13)C(C)C
95	TRUE	CCC(=O)OC(Cc1ccccc1)(C(C)CN(C)C)c2ccccc2
95	TRUE	CC(C)NCC(O)COc1cccc2ccccc12
62.5	UNKNOWN	CCCc1cc(=O)[nH]c(=S)[nH]1
95	TRUE	CNCCCC1c2ccccc2C=Cc3ccccc13
4	FALSE	OC(=O)c1cc2ccccc2c(Cc3c(O)c(cc4ccccc34)C(=O)O)c1O
100	TRUE	NC(=O)c1cnccn1
100	TRUE	CCCN1CC(CC2Cc3c(O)cccc3CC21)NS(=O)(=O)N(CC)CC
60	UNKNOWN	CCOC(=O)C(CCc1ccccc1)NC(C)C(=O)N2Cc3ccccc3CC2C(=O)O
63	UNKNOWN	Oc1ccc(cc1)c2sc3cc(O)ccc3c2C(=O)c4ccc(OCCN5CCCCC5)cc4
60	UNKNOWN	CCOC(=O)C(CCc1ccccc1)NC(C)C(=O)N2C(CC3CCCC32)C(=O)O
50	UNKNOWN	CNC(=C[N+](=O)[O-])NCCSCc1ccc(CN(C)C)o1
80	TRUE	CC(=CCO)C=CC=C(C)C=CC1=C(C)CCCC1(C)C
50	UNKNOWN	COC1C=COC2(C)Oc3c(C2=O)c4C5=NC6(CCN(CC(C)C)CC6)NC5=C(NC(=O)C(=CC=CC(C)C(O)C(C)C(O)C(C)C(OC(=O)C)C1C)C)C(=O)c4c(O)c3C
60	UNKNOWN	Nc1nc2ccc(OC(F)(F)F)cc2s1
97	TRUE	Cc1nc2CCCCn2c(=O)c1CCN3CCC(CC3)c4noc5cc(F)ccc45
95	TRUE	CC(NCCc1ccc(O)cc1)C(O)c2ccc(O)cc2
95	TRUE	CCCN(CCC)CCc1cccc2NC(=O)Cc21
75	UNKNOWN	OC(=O)c1ccccc1O
30	FALSE	CC(C)(C)NC(=O)C1CC2CCCCC2CN1CC(O)C(Cc3ccccc3)NC(=O)C(CC(=O)N)NC(=O)c4ccc5ccccc5n4
10	FALSE	CCCC[N+]1(C)C2CC(CC1C3OC32)OC(=O)C(CO)c4ccccc4
100	TRUE	CC(Cc1ccccc1)N(C)CC#C
5	FALSE	OCC1OC(Oc2cccc3C(C4c5cccc(OC6OC(CO)C(O)C(O)C6O)c5C(=O)c7c(O)cc(cc47)C(=O)O)c8cccc(O)c8C(=O)c23)C(O)C(O)C1O
100	TRUE	CC(C)NCC(O)c1ccc(NS(=O)(=O)C)cc1
82	TRUE	Cc1cn(C2OC(CO)C=C2)c(=O)[nH]c1=O
0	FALSE	CN(N=O)C(=O)NC1C(O)OC(CO)C(O)C1O
100	TRUE	Nc1ccc(cc1)S(=O)(=O)Nc2ncccn2
95	TRUE	Cc1cc(C)nc(NS(=O)(=O)c2ccc(N)cc2)n1
85	TRUE	Cc1cc(NS(=O)(=O)c2ccc(N)cc2)no1
25	FALSE	OC(=O)c1cc(N=Nc2ccc(cc2)S(=O)(=O)Nc3ccccn3)ccc1O
90	TRUE	CC1=C(CC(=O)O)c2cc(F)ccc2C1=Cc3ccc(cc3)S(=O)C
32.5	UNKNOWN	CCN1CCCC1CNC(=O)c2cc(ccc2OC)S(=O)(=O)N
100	TRUE	CNS(=O)(=O)Cc1ccc2[nH]cc(CCN(C)C)c2c1
0	FALSE	Cc1ccc(cc1NC(=O)c2cccc(NC(=O)Nc3cccc(c3)C(=O)Nc4cc(ccc4C)C(=O)Nc5ccc(c6cc(cc(c56)S(=O)(=O)O)S(=O)(=O)O)S(=O)(=O)O)c2)C(=O)Nc7ccc(c8cc(cc(c78)S(=O)(=O)O)S(=O)(=O)O)S(=O)(=O)O
95	TRUE	Nc1c2CCCCc2nc3ccccc13
95	TRUE	CN1C(=O)C(O)N=C(c2ccccc2)c3cc(Cl)ccc13
100	TRUE	CN1C(C(=O)Nc2ccccn2)C(=O)c3sccc3S1(=O)=O
95	TRUE	COc1cc2nc(nc(N)c2cc1OC)N3CCN(CC3)C(=O)C4CCCO4
80	TRUE	CN(CC=CC#CC(C)(C)C)Cc1cccc2ccccc12
100	TRUE	CC(C)(C)c1ccc(cc1)C(O)CCCN2CCC(CC2)C(O)(c3ccccc3)c4ccccc4

95	TRUE	COc1cc2CCN3CC(CC(C)C)C(=O)CC3c2cc1OC
90	TRUE	c1nc(cs1)c2nc3ccccc3[nH]2
34	UNKNOWN	Nc1nc2[nH]cnc2c(=S)[nH]1
60	UNKNOWN	CSc1ccc2Sc3ccccc3N(CCC4CCCCN4C)c2c1
90	TRUE	CC1CC2=C(CCC(=O)C2)C3CCC4(C)C(CCC4(O)C#C)C13
0	FALSE	CC1(C)SC2C(NC(=O)C(C(=O)O)c3ccsc3)C(=O)N2C1C(=O)O
80	TRUE	Clc1ccccc1CN2CCc3sccc3C2
100	TRUE	CC(C)(C)NCC(O)COc1nsnc1N2CCOCC2
100	TRUE	CCS(=O)(=O)CCn1c(C)ncc1[N+](=O)[O-]
0	FALSE	NCC1OC(OC2C(N)CC(N)C(OC3OC(CO)C(O)C(N)C3O)C2O)C(N)CC1O
100	TRUE	CC(N)C(=O)Nc1c(C)cccc1C
90	TRUE	C(C1=NCCN1)c2ccccc2
95	TRUE	CCCCNC(=O)NS(=O)(=O)c1ccc(C)cc1
99	TRUE	Cc1ccc(cc1)C(=O)c2ccc(CC(=O)O)n2C
66	UNKNOWN	COc1ccc2c(cccc2c1C(F)(F)F)C(=S)N(C)CC(=O)O
50	UNKNOWN	NCC1CCC(CC1)C(=O)O
100	TRUE	Clc1cccc(c1)N2CCN(CCCn3nc4ccccn4c3=O)CC2
85	TRUE	Cc1nnc2CN=C(c3ccccc3Cl)c4cc(Cl)ccc4-n12
100	TRUE	CN1CCN(CCCN2c3ccccc3Sc4ccc(cc24)C(F)(F)F)CC1
100	TRUE	OC(CCN1CCCCC1)(C2CCCCC2)c3ccccc3
95	TRUE	COc1cc(Cc2cnc(N)nc2N)cc(OC)c1OC
80	TRUE	CC(CN(C)C)CN1c2ccccc2CCc3ccccc13
53	UNKNOWN	Cc1c(C)c2OC(C)(COc3ccc(CC4SC(=O)NC4=O)cc3)CCc2c(C)c1O
95	TRUE	CCCC(CCC)C(=O)O
0	FALSE	CNC(CC(C)C)C(=O)NC1C(O)c2ccc(Oc3cc4cc(Oc5ccc(cc5Cl)C(O)C6NC(=O)C(NC(=O)C4NC(=O)C(CC(=O)N)NC1=O)c7ccc(O)c(c7)-c8c(O)cc(O)cc8C(NC6=O)C(=O)O)c3OC9OC(CO)C(O)C(O)C9OC%10CC(C)(N)C(O)C(C)O%10)c(Cl)c2
92	TRUE	COc1ccc(cc1)C(CN(C)C)C2(O)CCCCC2
90	TRUE	COc1ccc(CCN(C)CCCC(C#N)(C(C)C)c2ccc(OC)c(OC)c2)cc1OC
80	TRUE	NC(CCC(=O)O)C=C
95	TRUE	CC(=O)CC(C1C(=O)Oc2ccccc2C1=O)c3ccccc3
73	UNKNOWN	Cc1cccc(C)c1NC(=O)c2cc(c(Cl)cc2O)S(=O)(=O)N
85	TRUE	Nc1ccn(C2CCC(CO)O2)c(=O)n1
95	TRUE	CN(C)C(=O)Cc1c(nc2ccc(C)cn12)c3ccc(C)cc3
95	TRUE	CN1CCN(CC1)C(=O)OC2N(C(=O)c3nccnc23)c4ccc(Cl)cn4
95	TRUE	CN(C)CCOC1=Cc2ccccc2Sc3ccc(Cl)cc13
92	TRUE	CC1=NN=C2N1C3=C(C=C(C=C3)Cl)C(=NC2)C4=CC=CC=C4
90	TRUE	CC1CC2C3CCC4=CC(=O)C=CC4(C3(C(CC2(C1(C(=O)CO)O)C)O)F)C
100	TRUE	C(C1C(C(C(C(O1)O)O)O)O)O
95	TRUE	CCN(CC)C(C)C(=O)C1=CC=CC=C1
10	FALSE	C(CN(CC(=O)[O-])CC(=O)[O-])N(CC(=O)[O-])CC(=O)[O-]
100	TRUE	CCO
100	TRUE	CC12CCC3C(C1CCC2(C#C)O)CCC4=C3C=CC(=C4)O
100	TRUE	C(C(C(C(C(=O)CO)O)O)O)O
100	TRUE	CC12CCC(=O)C=C1CCC3C2C(CC4(C3CCC4(C(=O)CO)O)C)O
0	FALSE	CC(=O)N(CC(CO)O)C1=C(C(=C(C(=C1I)C(=O)NCC(CO)O)I)C(=O)NCC(CO)O)I
0	FALSE	C1=C(C(=C(C(=C1I)I)NC(=O)COCCOCCOCC(=O)NC2=C(C(=CC(=C2I)I)I)C(=O)O)C(=O)O)I
90	TRUE	C1C(C2C(O1)C(CO2)O[N+](=O)[O-])O[N+](=O)[O-]
100	TRUE	C1C(C2C(O1)C(CO2)O[N+](=O)[O-])O
0	FALSE	CCNC(=O)C1CCCN1C(=O)C(CCCN=C(N)N)NC(=O)C(C(C)C)NC(=O)C(C(C)C)NC(=O)C(CC2=CC=C(C=C2)O)NC(=O)C(CO)NC(=O)C(CC3=CNC4=CC=CC=C43)NC(=O)C(CC5=CN=CN5)NC(=

		O)C6CCC(=O)N6
10	FALSE	C[N+](C)(C)CC(CC(=O)[O-])O
100	TRUE	CCC12CCC3C(C1CCC2(C#C)O)CCC4=CC(=O)CCC34
57	UNKNOWN	C1=C(C=C(C(=C1I)OC2=CC(=C(C(=C2)I)O)I)I)CC(C(=O)O)N
90	TRUE	C1CC(=C(N2C1C(C2=O)NC(=O)C(C3=CC=CC=C3)N)C(=O)O)Cl
76	UNKNOWN	C(CS(=O)(=O)[O-])S
90	TRUE	C1=CC(=CN=C1)C(=O)[O-]
0	FALSE	CC(C1C(=O)NC(CSSCC(C(=O)NC(C(=O)NC(C(=O)NC(C(=O)N1)CCCCN)CC2=CNC3=CC=CC=C32)CC4=CC=CC=C4)NC(=O)C(CC5=CC=CC=C5)N)C(=O)NC(CO)C(C)O)O
90	TRUE	CCCC1=NC=CC(=C1)C(=S)N
95	TRUE	CC(C(C1=CC=CC=C1)O)NC
15	FALSE	C[N+]1=CC=CC(=C1)OC(=O)N(C)C
80	TRUE	COC1=CC2=C(C=CN=C2C=C1)C(C3CC4CCN3CC4C=C)O
30	FALSE	COC1C(CC2CN3CCC4=C(C3CC2C1C(=O)OC)NC5=C4C=CC(=C5)OC)OC(=O)C6=CC(=C(C(=C6)OC)OC)OC
45	UNKNOWN	C1=NC(=NN1C2C(C(C(O2)CO)O)O)C(=O)N
1	FALSE	C(C1C(C(C(C(O1)OC2(C(C(C(O2)OS(=O)(=O)[O-])OS(=O)(=O)[O-])OS(=O)(=O)[O-])COS(=O)(=O)[O-])OS(=O)(=O)[O-])OS(=O)(=O)[O-])OS(=O)(=O)[O-])OS(=O)(=O)[O-]
2	FALSE	C1C2C(=O)NC(C3=CC(=CC(=C3)O)OC4=C(C=CC(=C4)C(C(=O)N2)N)O)C(=O)NC5C6=CC(=C(C(=C6)OC7=C(C=C(C=C7)C(C8C(=O)NC(C9=CC(=CC(=C9C2=C(C=CC(=C2)C(C(=O)N8)NC5=O)O)O)O)C(=O)O)O)Cl)O)OC2=C(C=C1C=C2)Cl
0	FALSE	C1CC(N(C1)C(=O)C2CSSCC(C(=O)NC(C(=O)NC(C(=O)NC(C(=O)NC(C(=O)N2)CC(=O)N)CCC(=O)N)CC3=CC=CC=C3)CC4=CC=C(C=C4)O)NC(=O)CNC(=O)CNC(=O)CN)C(=O)NC(CCCCN)C(=O)NCC(=O)N
95	TRUE	CN1C2=C(C(=O)N(C1=O)C)NC=N2
80	TRUE	CC(CN1C2=CC=CC=C2SC3=CC=CC=C31)CN(C)C
100	TRUE	CC(CCC(=O)O)C1CCC2C1(CCC3C2C(CC4C3(CCC(C4)O)C)O)C
0	FALSE	C1=NC2=C(C(=N1)N)N=CN2C3C(C(C(O3)CO)O)O
70	UNKNOWN	CC1=C2C(=C(C(=C1C)OC(=O)C)C)CCC(O2)(C)CCCC(C)CCCC(C)CCCC(C)C
95	TRUE	C1CN(CCN1CCC=C2C3=CC=CC=C3SC4=C2C=C(C=C4)Cl)CCO

7.4 Supporting information Project 5.5

CID	Outcome	SMILES
24867529	Active	COC1=CC2=C(C=CN=C2C=C1)[C@H]([C@H]3C[C@H]4CCN3C[C@H]4C=C)O
11957564	Active	C1CCC(CC1)[Si](CCCN2CCCCC2)(C3=CC=C(C=C3)F)O
3068143	Active	CC[C@H]1CN2CCC3=CC(=C(C=C3[C@@H]2C[C@@H]1C[C@@H]4C5=CC(=C(C=C5CCN4)OC)OC)OC)OC
24867531	Active	COC1=CC2=C(C=CN=C2C=C1)[C@H]([C@@H]3C[C@H]4CCN3C[C@H]4C=C)O
2733504	Active	C1=CC=C2C(=C1)C3=CC=CC=C3[I+]2
6239	Active	CCN(CC)CCCC(C)NC1=C2C=C(C=CC2=NC3=C1C=CC(=C3)Cl)OC
11957453	Active	C[N+](C)(CCC(=O)CC[N+](C)(C)C1=CC=C(C=C1)CC=C)C2=CC=C(C=C2)CC=C
6604151	Active	C1=CC(=CC=C1C(=N)N)OCCCCCOC2=CC=C(C=C2)C(=N)N
9853645	Active	CC1=[N+](C2=CC=CC=C2C(=C1)N)CCCCCCCCCCCCCC[N+]3=C(C=C(C4=CC=CC=C43)N)C
11957525	Active	C1=CC(=C(C=C1CN=C(N)NC(=O)C2=C(N=C(C(=N2)Cl)N)N)Cl)Cl
10440396	Active	C[N+](C)(CCCCCC[N+](C)(C)CCCN1C(=O)C2=CC=CC3=C2C(=CC=C3)C1=O)CCCN4C(=O)C5=CC=CC6=C5C(=CC=C6)C4=O
824226	Active	C1=CC(=CC=C1C2=NC3=C(N2)C=C(C=C3)C4=NC5=C(N4)C=C(C=C5)N)N
184822	Active	CCCNC[C@@H](COC1=CC=CC=C1C(=O)CCC2=CC=CC=C2)O
60703	Active	C1CN(CCC1CC2=CC=C(C=C2)F)CC(C3=CC=C(C=C3)Cl)O
36708	Active	CCCNCC(COC1=CC=CC=C1C(=O)CCC2=CC=CC=C2)O
2812	Active	C1=CC=C(C=C1)C(C2=CC=CC=C2)(C3=CC=CC=C3Cl)N4C=CN=C4
24867458	Active	CC(C)[C@@]1(C(=O)N2[C@H](C(=O)N3CCCC3[C@@]2(O1)O)CC4=CC=CC=C4)NC(=O)[C@@H]5C[C@H]6[C@@H](CC7=CNC8=CC=CC6=C78)N(C5)C
11957606	Active	CC[N+](CC)(CC)COC1=CC=C(C=C1)/C=C/C2=CC=CC=C2

5326739	Active	C1=CC=C2C(=C1)C(=C(N2)C3=C4C=CC=CC4=NC3=O)NO
5280754	Active	CCC1C(=O)N(CC(=O)N(C(C(=O)NC(C(=O)N(C(C(=O)NC(C(=O)NC(C(=O)N(C(C(=O)N(C(C(=O)N(C(C(=O)N(C(C(=O)N1)C(C(C)C/C=C/C)O)C)C(C)C)C)CC(C)C)C)CC(C)C)C)C)C)CC(C)C)C)C(C)C)CC(C)C)C)C
126941	Active	CN(CC1=CN=C2C(=N1)C(=NC(=N2)N)N)C3=CC=C(C=C3)C(=O)N[C@@H](CCC(=O)O)C(=O)O
107656	Active	CC1=C(C(=C2CCC(OC2=C1C)(C)CN3CCN(CC3)C4=NC(=NC(=C4)N5CCCC5)N6CCCC6)C)O
51082	Active	C1=CC(=C2C(=C1NCCNCCO)C(=O)C3=C(C=CC(=C3C2=O)O)O)NCCNCCO
5614	Active	CC(C)(C)C1=CC(=CC(=C1O)C(C)(C)C)C=C(C#N)C#N
4477	Active	C1=CC(=C(C=C1[N+](=O)[O-])Cl)NC(=O)C2=C(C=CC(=C2)Cl)O
3213	Active	CC1=C2C(=C(C3=C1C=CN=C3)C)C4=CC=CC=C4N2
11957726	Active	CCCN(CCC)[C@@H]1CCC2=C(C=CC(=C2C1)O)F
11957700	Active	CCC1=CC=CC=C1OC[C@H](CN[C@H]2CCC3=CC=CC=C3C2)O
11957587	Active	CC[C@@H](C)[C@@H]1[C@H](CC[C@@]2(O1)C[C@@H]3C[C@H](O2)C/C=C(/[C@H]([C@H](/C=C/C=C/4\CO[C@H]5[C@@]4([C@@H](C=C([C@H]5O)C)C(=O)O3)O)C)O[C@H]6C[C@@H]([C@H]([C@@H](O6)C)O[C@H]7C[C@@H]([C@H]([C@@H](O7)C)O)OC)OC)\C)C
11957499	Active	C[C@@H]1CC[C@]2([C@@H](C[C@H]([C@H](O2)[C@H](C)C(=O)C3=CC=CN3)C)C)O[C@@H]1CC4=NC5=C(O4)C=CC(=C5C(=O)O)NC
10236521	Active	CCCN(CCC)[C@H]1CCC2=C(C=CC(=C2C1)O)F
4605800	Active	C1CC2=C3C(=CC=C2)N(S(=O)(=O)N3C1)CCN4CCC(=CC4)C5=CNC6=C5C=CC(=C6)F
42890	Active	C[C@H]1[C@H]([C@H](C[C@@H](O1)O[C@H]2C[C@@](CC3=C(C4=C(C(=C23)O)C(=O)C5=CC=CC=C5C4=O)O)(C(=O)C)O)N)O
24195776	Active	CCCN1CCC2=CC=CC3=C2C1CC4=C3C(=C(C=C4)O)O
11957469	Active	C[C@H]1CCC/C=C/C2C[C@@H](C[C@]2(C/C=C/C(=O)O1)O)O
5329255	Active	C1=CC(=C(C=C1/C=C(/C(=O)NCCCNC(=O)/C(=C/C2=CC(=C(C=C2)O)O)/C#N)\C#N)O)O
1318	Active	C1=CC2=C(C3=C(C=CC=N3)C=C2)N=C1
73334	Active	CCNC(=O)N1CCN(CC1)CCCC(C2=CC=C(C=C2)F)C3=CC=C(C=C3)F
24867476	Active	CC(C)C[C@H]1C(=O)N2CCCC2[C@]3(N1C(=O)[C@](O3)(C(C)C)NC(=O)[C@H]4CN([C@@H]5CC6=C(NC7=CC=CC(=C67)C5=C4)Br)C)O
148673	Active	COC1=C(C=CC(=C1)NS(=O)(=O)C)NC2=C3C=CC=CC3=NC4=CC=CC=C42
5770	Active	CO[C@H]1[C@@H](C[C@@H]2CN3CCC4=C([C@H]3C[C@@H]2[C@@H]1C(=O)OC)NC5=C4C=CC(=C5)OC)OC(=O)C6=CC(=C(C(=C6)OC)OC)OC
24867499	Active	C=CCN1CCC2=CC=CC3=C2C1CC4=C3C(=C(C=C4)O)O
24867538	Active	C[C@]12CC=C3C([C@@H]1CC[C@@H]2C(=O)CN4CCN(CC4)C5=NC(=NC(=C5)N6CCCC6)N7CCCC7)CCC8=CC(=O)C=C[C@@]83C
11957693	Active	COC1=CC=C(C=C1)CCCOC2=C(C=CC(=C2)CCN3C=CN=C3)OC
11957671	Active	CCCN(CCC1=CC=CC=C1)C2CCC3=C(C2)C=CC=C3O
443390	Active	COC1=C(C=C2C(=C1)CCN2C(=O)NC3=CC(=CC(=C3)C4=CN=CC=C4)F)C(F)(F)F
441276	Active	CC1=C2[C@H](C(=O)[C@@]3([C@H](C[C@@H]4[C@](C3[C@@H]([C@@](C2(C)C)(C[C@@H]1OC(=O)[C@@H]([C@H](C5=CC=CC=C5)NC(=O)C6=CC=CC=C6)O)O)OC(=O)C7=CC=CC=C7)(CO4)OC(=O)C)O)C)OC(=O)C
107759	Active	COC1=CC=CC=C1CNCCCCCCNCCCCCCCCNCCCCCCNCC2=CC=CC=C2OC
64927	Active	CCN(CC)CCCC(C)NC1=C2C=CC(=CC2=NC=C1)Cl
65341	Active	CCCC(C1=CC=CC=C1)(C2=CC=CC=C2)C(=O)OCCN(CC)CC
122215	Active	CCCCCCCCCC[Si](C)(C)CCC(=O)NC(CC1=CC=C(C=C1)C)C2=CC=CC=C2
16759248	Active	CCCCN1C2CCC1CC(C2)OC(C3=CC=C(C=C3)F)C4=CC=C(C=C4)F
5702010	Active	CN1CCC2=CC=CC3=C2C1CC4=C3C(=C(C=C4)O)O
9951033	Active	C1[C@H](O[C@H](C2=C1C(=C(C=C2)O)O)CN)C34CC5CC(C3)CC(C5)C4
10649	Active	CC1=[N+](C2=CC=CC=C2C(=C1)N)CCCCCCCCCC[N+]3=C(C=C(C4=CC=CC=C43)N)C
24867491	Active	C[C@@]1(C(=O)N2[C@H](C(=O)N3CCCC3[C@@]2(O1)O)CC4=CC=CC=C4)NC(=O)[C@@H]5C[C@H]6[C@@H](CC7=CNC8=CC=CC6=C78)N(C5)C
41114	Active	CC1=C(C(C(=C(N1)C)C(=O)OCCN(C)CC2=CC=CC=C2)C3=CC(=CC=C3)[N+](=O)[O-])C(=O)OC
10047903	Active	CC(COC1=CC=C(C=C1)/C=C/C2=CC=CC=C2)[N+](C)(C)C
441325	Active	CCCCC1=C(C2=CC=CC=C2O1)C(=O)C3=CC(=C(C(=C3)I)OCCN(CC)CC)I
62978	Active	COC1=C(C=C2C(=C1)C(=NC(=N2)N3CCN(CC3)C(=O)C4COC5=CC=CC=C5O4)N)OC
11957656	Active	C=CCN1C2[C@]3([C@]4(C5=C(C2)C=CC(=C5O[C@@H]4/C(=N/N=C\6/[C@H]7OC8=C(C=CC9=C8[C@]72[C@](C(C9)N(CC2)CC=C)(CC6)O)O)/CC3)O)CC1)O
24360	Active	CC[C@@]1(C2=C(COC1=O)C(=O)N3CC4=CC5=CC=CC=C5N=C4C3=C2)O
9874535	Active	COC1=CC(=CC(=C1OC)OC)C2=C(N(C(=O)C3=C2C=CC(=C3)OCC4=CC=CC=N4)C5=CC=C(C

		=C5)N)C(=O)OC
5702295	Active	C1=CC=C(C=C1)CN=C(N)NC(=O)C2=C(N=C(C(=N2)Cl)N)N
5287844	Active	C1=CC2=C(/C(=C/3\C4=C(C=C(C=C4)Br)NC3=O)/N=C2C=C1)NO
71420	Active	CN(C)C(=O)C(CCN1CCC(CC1)(C2=CC=C(C=C2)Cl)O)(C3=CC=CC=C3)C4=CC=CC=C4
68635	Active	C[N+]1=C2C(=C3C=CC4=C(C3=C1)OCO4)C=CC5=CC6=C(C=C52)OCO6
54900	Active	C1CCN(CC1)CCOC2=CC=C(C=C2)C(=O)C3=C(SC4=C3C=CC(=C4)O)C5=CC=C(C=C5)O
443600	Active	CCCCC[C@H](CC(=O)NO)C(=O)N[C@@H](C(C)C)C(=O)N1CCC[C@H]1CO
11957578	Active	CC(C)C1=C2C[C@](CCC2=CC(=C1)F)(CCN(C)CCCC3=NC4=CC=CC=C4N3)OC(=O)C5CC5
6435335	Active	CCOC(=O)NC1=C(N=C(C=C1)NCC2=CC=C(C=C2)F)N
11957719	Active	CC[C@@]12C=CCN3[C@@H]1[C@]4(CC3)C([C@]([C@@H]2OC(=O)C)(C(=O)OC)O)N(C5=CC(=C(C=C45)[C@]6(CC7CC(CN(C7)CCC8=C6NC9=CC=CC=C89)(CC)O)C(=O)OC)OC)C
11957677	Active	CCCCCCCCCC(=O)NC(CN1CCOCC1)[C@@H](C2=CC=CC=C2)O
104895	Active	CCCCCCC(C)(C)C1=CC(=C(C=C1)[C@@H]2C[C@@H](CC[C@H]2CCCO)O)O
11957588	Active	C1=CC(=CC=C1C(C2=CC=C(C=C2)F)OCCCC3=CN=CN3)F
1730	Active	C1=CC(=CC=C1C(=O)O)[Hg]Cl
24867482	Active	C[C@]1(CCCC(C1)C(C)(C)NC(=O)CBr)NCC(COC2=CC=CC=C2CC=C)O
5282483	Active	C1CN(CCN1CC/C=C\2/C3=CC=CC=C3SC4=C2C=C(C=C4)C(F)(F)F)CCO
3060974	Active	CC1=CC(=C(C(=C1)C)N=C2C=C3C4=CC(=C(C=C4CCN3C(=O)N2C)OC)OC)C
203135	Active	CN1C2CCC1CC(C2)OC(C3=CC=CC=C3)C4=CC=C(C=C4)Cl
173603	Active	CCCN1CCC2=C3C1CC4=C(C3=CC(=C2)O)C(=C(C=C4)O)O
60662	Active	CC(C)[C@H]1C2=C(CC[C@@]1(CCN(C)CCCC3=NC4=CC=CC=C4N3)OC(=O)COC)C=C(C=C2)F
11957481	Active	CC(C)C[C@@H](C(=O)O)NC(=O)[C@H]([C@@H](CC1=CC=CC=C1)N)O
5280343	Active	C1=CC(=C(C=C1C2=C(C(=O)C3=C(C=C(C=C3O2)O)O)O)O)O
10921	Active	C[N+](C)(C)CCCCCCCCCC[N+](C)(C)C
2544	Active	CC1(C2CCC(C1(C)C(=O)O)O2)C(=O)O
10321498	Active	CCN(CC)C(=O)C1=CC=C(C=C1)[C@H](C2=CC(=CC=C2)OC)N3C[C@@H](N(C[C@H]3C)CC=C)C
5281847	Active	CC1=C(C(=C(C(=C1O)C(=O)C)O)CC2=C(C3=C(C(=C2O)C(=O)/C=C/C4=CC=CC=C4)OC(C=C3)(C)C)O)O
176157	Active	CCCN(CC1CC1)C2=NC(=NC(=C2Cl)NC3=C(C=C(C=C3Cl)Cl)Cl)C
11497466	Active	CC1=C(C=CC(=C1)C2=NOC(=N2)C)C3=CC=C(C=C3)C(=O)NC4=CC(=C(C=C4)OC)N5CCN(CC5)C
9951825	Active	C1=CC=C(C(=C1)N)S/C(=C(\C(=C(/SC2=CC=CC=C2N)\N)\C#N)/C#N)/N
3108	Active	C1CCN(CC1)C2=NC(=NC3=C2N=C(N=C3N4CCCCC4)N(CCO)CCO)N(CCO)CCO
2545	Active	CC12C3CCC(C1(C(=O)OC2=O)C)O3
24867460	Active	CCCC1O[C@@H]2C[C@H]3[C@@H]4CCC5=CC(=O)C=C[C@@]5(C4[C@H](C[C@@]3([C@@]2(O1)C(=O)CO)C)O)C
11957716	Active	C[C@]12CCC3C([C@@H]1CC[C@]2(C#N)O)CCC4=CC5=NC6=NC7=CC=CC=C7N6C=C5C[C@]34C
11957542	Active	CC(C)C[C@@H](C(=O)NCCCCN=C(N)N)NC(=O)C1C(O1)C(=O)O
5283454	Active	CCCCCCCC/C=C\CCCCCCCC(=O)NCCO
24867497	Active	C[C@@H]1C2[C@@H]([C@H]3C(C(=O)/C(=C(/N)\O)/C(=O)[C@]3(C(=O)C2=C(C4=C1C=CC=C4O)O)O)N(C)C)O
10098248	Active	CC(C)(C)NC(=O)C1CN(CCN1C(=O)OCC2=CC=CC=C2)C3=NC4=CC(=C(C=C4C(=N3)N)OC)OC
5405	Active	CC(C)(C)C1=CC=C(C=C1)C(CCCN2CCC(CC2)C(C3=CC=CC=C3)(C4=CC=CC=C4)O)O
4748	Active	C1CN(CCN1CCCN2C3=CC=CC=C3SC4=C2C=C(C=C4)Cl)CCO
3396	Active	C1CN(CCC12C(=O)NCN2C3=CC=CC=C3)CCCC(C4=CC=C(C=C4)F)C5=CC=C(C=C5)F
24867511	Active	CC(C)C1[C@@H]2C[C@@H]3CC4=C(C=CC(=C4C(=C3C(=O)[C@@]2(C(=O)/C(=C(/N)\O)/C1=O)O)O)O)N(C)C
9852041	Active	CC1=NC=CC(=C1)CN2C(=C(C3=C(C2=O)C(=NC=C3)OCC4=NC=CC=N4)C5=CC(=C(C(=C5)OC)OC)OC)C(=O)OC
67356	Active	C1CN(CCN1CCCN2C3=CC=CC=C3SC4=C2C=C(C=C4)C(F)(F)F)CCO
29976	Active	COC1=CC(=CC(=C1OC)OC)C(=O)OCCC[NH+]2CCC[NH+](CC2)CCCOC(=O)C3=CC(=C(C(=C3)OC)OC)OC
24867479	Active	CC(C)(C)[C@@]1(CCN2CC3C4=CC=CC=C4CCC5=C3C(=CC=C5)[C@@H]2C1)O
16362	Active	C1CN(CCC1N2C3=CC=CC=C3NC2=O)CCCC(C4=CC=C(C=C4)F)C5=CC=C(C=C5)F

4712	Active	C1=CC(=CC=C1C2=NC(=C(N2)C3=CC=NC=C3)C4=CC=C(C=C4)F)[N+](=O)[O-]
1967	Active	CC1=C(C(=O)C(=C(C1=O)C)CCCCC#CCCCC#CCO)C
11957714	Active	CC[C@@]12C=CCN3[C@@H]1[C@]4(CC3)C([C@]([C@@H]2OC(=O)C)(C(=O)OC)O)N(C5=CC(=C(C=C45)[C@]6(CC7CC(CN(C7)CCC8=C6NC9=CC=CC=C89)(CC)O)C(=O)OC)OC)C=O
11957662	Active	C1=CC(=CC(=C1)CCSC(=N)N)CCSC(=N)N
11957527	Active	C[C@@H]1OC[C@@H]2[C@@](O1)(C[C@H]([C@@H](O2)O[C@H]3[C@H]4COC(=O)[C@@H]4[C@@H](C5=CC6=C(C=C35)OCO6)C7=CC(=C(C(=C7)OC)O)OC)O)O
11647992	Active	C1CN(CCC1(C2=CC(=CC=C2)C(F)(F)F)O)CCCC(=O)C3=CC=C(C=C3)F
3559	Active	C1CN(CCC1(C2=CC=C(C=C2)Cl)O)CCCC(=O)C3=CC=C(C=C3)F
5312137	Active	CN(C)S(=O)(=O)C1=CC\2=C(C=C1)NC(=O)/C2=C\C3=CC4=C(N3)CCCC4
262093	Active	C1=CC=C2C(=C1)C(=O)C(=C(C2=O)SCCO)SCCO
11957570	Active	CCCN(CCC)[C@@H]1CCC2=C(C1)C(=CC=C2)O
11957495	Active	CCN1C=NC2=C1N=C(N=C2NC3=CC(=CC=C3)Cl)N[C@@H]4CCCC[C@@H]4N
6603857	Active	CN(C)C1=NC=C2C(=C1)C(=NC=N2)NC3=CC4=C(C=C3)N(N=C4)CC5=CC=CC=C5
5282407	Active	C1CN(CCN1C/C=C/C2=CC=CC=C2)C(C3=CC=C(C=C3)F)C4=CC=C(C=C4)F
66368	Active	CC(C)NCC(COC1=CC=CC=C1CC=C)O
9868848	Active	CC(C1=CC=CC=C1)(C2=CC=C(C=C2)Cl)OCCC3CCCN3C
3038495	Active	COC1=CC=CC=C1N2CCN(CC2)CCCCNC(=O)C3=CC4=CC=CC=C4C=C3
517348	Active	C1CCN(C1)C(=S)[S-]
104920	Active	C1CN(CCN1CCCC2=CC=CC=C2)CCOC(C3=CC=C(C=C3)F)C4=CC=C(C=C4)F
72430	Active	CC(C)[C@@H](C(=O)N[C@@H](CC1=CC=CC=C1)C=O)NC(=O)OCC2=CC=CC=C2
62969	Active	CC(C)C(CCCN(C)CCC1=CC(=C(C=C1)OC)OC)(C#N)C2=CC(=C(C=C2)OC)OC
5546	Active	C1=CC=C(C=C1)C2=NC3=C(N=C2N)N=C(N=C3N)N
11957577	Active	C[C@@H](CN1CCC2=C1C=C(C=C2)Br)N
644274	Active	C1=CC(=CC=C1C(C2=CC=C(C=C2)Cl)N3C=C[N+](=C3)CC(C4=C(C=C(C=C4)Cl)Cl)OCC5=C(C=C(C=C5)Cl)Cl)Cl
3151	Active	C1CN(CCC1N2C3=C(C=C(C=C3)Cl)NC2=O)CCCN4C5=CC=CC=C5NC4=O
11957697	Active	CN1CCC2=C(C(=C(C=C2C(C1)C3=CC(=CC=C3)Cl)O)O)Cl
6603792	Active	C1CCC(C1)N2C=C(C3=C2N=CN=C3N)C4=CC=C(C=C4)OC5=CC=CC=C5
5487525	Active	CC(C)(C)C1=CC(=C/C(=C(/N)\S)/C#N)C=C(C1=O)C(C)(C)C
135348	Active	C1CN(CCC1C(=O)C2=CC=C(C=C2)F)CCN3C(=O)C4=CC=CC=C4NC3=O
132496	Active	CCCCCCN(CCCCCC)C(=O)CC1=C(NC2=CC=CC=C21)C3=CC=C(C=C3)F
119442	Active	CC(C)C(CCCN(C)CCC1=CC(=C(C=C1)OC)OC)(C#N)C2=CC(=C(C(=C2)OC)OC)OC
91505	Active	CCN(CC)CC#CCOC(=O)C(C1CCCCC1)(C2=CC=CC=C2)O
66366	Active	CC(C)NC[C@H](COC1=CC=CC2=CC=CC=C21)O
3957	Active	CCOC(=O)N1CCC(=C2C3=C(CCC4=C2N=CC=C4)C=C(C=C3)Cl)CC1
24867500	Active	CN(C)C1[C@@H]2CC3[C@@H](C4=C(C=CC(=C4C(=C3C(=O)[C@@]2(C(=O)/C(=C(\N)/O)/C1=O)O)O)O)Cl)O
11957722	Active	CCN(CC)C(=O)N[C@@H]1C[C@H]2[C@@H](CC3=CNC4=CC=CC2=C34)N(C1)C
11957605	Active	C1CN2CCC1C(C2C(C3=CC=CC=C3)C4=CC=CC=C4)NCC5=CC=CC=C5I
5282408	Active	CN1CCC(=C2C3=C(C(=O)CC4=CC=CC=C42)SC=C3)CC1
71587	Active	C1CN(CCN1CCCN2C3=CC=CC=C3C=CC4=CC=CC=C42)CCO
11957579	Active	CC(C(C1=CC=C(C=C1)O)O)N2CCC(CC2)CC3=CC=CC=C3
6604029	Active	C1=C(C=C(C(=C1O)O)O)/C(=C(\C=C(C#N)C#N)/N)/C#N
5702062	Active	CC(=O)O
62882	Active	CC(C)NCC(COC1=CC=CC2=CC=CC=C21)O
6014	Active	CC(CN1C2=CC=CC=C2SC3=CC=CC=C31)N(C)C
2051	Active	COC1=C(C=C2C(=C1)C(=NC=N2)NC3=CC(=CC=C3)Cl)OC
11957725	Active	C1CCN(CC1)CC2=CC(=CC=C2)OCCCNC3=NC4=CC=CC=C4S3
11957483	Active	C=CCN1CCC2=C(C(=C(C=C2[C@H](C1)C3=CC=CC=C3)O)O)Br
9909521	Active	C=CCN1CCC2=C(C(=C(C=C2C(C1)C3=CC=CC=C3)O)O)Cl
4519262	Active	CC(C)C1=CC2=C(C=C1)N(C(=C2SC(C)(C)C)CC(C)(C)C(=O)[O-])CC3=CC=C(C=C3)Cl
238053	Active	CN1C2CCC1CC(C2)OC(C3=CC=CC=C3)C4=CC=CC=C4

71401	Active	C1=CC(=C(C(=C1)Cl)CC(=O)N=C(N)N)Cl
28693	Active	CN1C[C@@H](C[C@H]2[C@H]1CC3=CN(C4=CC=CC2=C34)C)CNC(=O)OCC5=CC=CC=C5
24867540	Active	CCCCCCCC(=O)O[C@@H]1[C@H](C(=C2C1[C@@](C[C@H]([C@]3([C@H]2OC(=O)[C@@]3(C)O)O)OC(=O)CCC)(C)OC(=O)C)C)OC(=O)/C(=C\C)/C
11957658	Active	CS(=O)(=O)O
11957596	Active	CN([C@@H]1CCCC[C@@H]1N2CCCC2)C(=O)CC3=CC(=C(C=C3)Cl)Cl
719408	Active	C1=CC=C(C=C1)NC(=S)NC2=NC=CS2
73314	Active	CC(C)CN(C)C1=NC(=C(N=C1Cl)C(=O)N=C(N)N)N
66064	Active	CN1CCN(CC1)CCCN2C3=CC=CC=C3SC4=C2C=C(C=C4)C(F)(F)F
6240	Active	CN(C)CCCN1C2=CC=CC=C2SC3=C1C=C(C=C3)Cl
5578	Active	COC1=CC(=CC(=C1OC)OC)CC2=CN=C(N=C2N)N
11957590	Active	CC1=C2CCCC2=C(C=C1)OCC(C(C)NC(C)C)O
11957508	Active	C1CN(C=CC1N2C3=CC=CC=C3NC2=O)CCCC(=O)C4=CC=C(C=C4)F
11957471	Active	CN(C)CC/C=C/1\C2=C(C=C(C=C2)Cl)SC3=CC=CC=C31
11417991	Active	C1CN(CCC1OC(=O)C(C2=CC=CC=C2)C3=CC=CC=C3)CCCl
9604979	Active	CS(=O)(=O)O
6604176	Active	CC1=C(C2=C3N1[C@@H](COC3=CC=C2)CN4CCOCC4)C(=O)C5=CC=CC6=CC=CC=C65
13770	Active	CN1CCC(=C2C3=CC=CC=C3C=CC4=CC=CC=C42)CC1
11957553	Active	C1CN(CCN1CCCC2=CC=CC=C2)CCOC(C3=CC=CC=C3)C4=CC=CC=C4
11538542	Active	CCOC(=O)C1=C(N(C(=C(C1C2=CC(=CC=C2)[N+](=O)[O-])C(=O)OC)C)CC#C)C
10008573	Active	CN1C=C(N=C1C2=CC=C(C=C2)OCC(CNCCOC3=CC(=C(C=C3)O)C(=O)N)O)C(F)(F)F
5289419	Active	COC1=CC\2=C(C=C1)NC(=O)/C2=C\C3=CN=CN3
5288209	Active	CC1=C(C(CCC1)(C)C)/C=C/C(=C/C=C/C(=C/C(=O)NC2=CC=C(C=C2)O)/C)/C
5281032	Active	CN1CCN(CC1)CCCN2C3=C(SC4=CC=CC=C24)C=CC(=C3)Cl
4350931	Active	CN1CCC2=CC=CC=C2CC3=C(CC1)C4=CC=CC=C4N3
133633	Active	C1CN(CCC1(C2=CC=C(C=C2)Cl)O)CC3=CNC4=CC=CC=C43
66062	Active	CN1CCCCC1CCN2C3=CC=CC=C3SC4=C2C=C(C=C4)SC
62878	Active	C1CNC[C@H]([C@@H]1C2=CC=C(C=C2)F)COC3=CC4=C(C=C3)OCO4
60839	Active	CCC1=CC(=CC=C1)N(C)C(=NC2=CC=CC3=CC=CC=C32)N
3885	Active	CC1(CCC2=C(O1)C3=CC=CC=C3C(=O)C2=O)C
24867502	Active	CN1CCC2=C3C1CC4=C(C3=CC(=C2)O)C(=C(C=C4)O)O
11957699	Active	C1CN(C2=CC=CC=C21)C(=O)C(CC3=CC(=C(C=C3)O)O)C#N
11957600	Active	C1C[C@@H]([C@@H](NC1)C2=CC=CC=C2)OCC3=CC(=CC(=C3)C(F)(F)F)C(F)(F)F
11957569	Active	C1CN=C(N1)N(CC2=CC=C(C=C2)F)C3=C(C=CC=C3Cl)Cl
11226716	Active	CC1=C(C=CC(=C1)C2=NOC(=N2)C)C3=CC=C(C=C3)C(=O)N4CCC5=C4C=C6C(=C5)OCC67CCN(CC7)C
10314472	Active	CN[C@@H]1C[C@H](C2=CC=CC=C12)C3=CC(=C(C=C3)Cl)Cl
9601084	Active	C=CCN1CC[C@]23[C@@H]4/C(=N/NC(=O)C5=CC=CC=C5)/CC[C@]2([C@H]1CC6=C3C(=C(C=C6)O)O4)O
6438352	Active	CN1CCN(CC1)C2=NC3=C(C=CC(=C3)C(F)(F)F)N4C2=CC=C4
3014059	Active	C[N+]1(CCC(CC1)OC(=O)C(C2=CC=CC=C2)C3=CC=CC=C3)C
2733525	Active	CC/C(=C(\C1=CC=CC=C1)/C2=CC=C(C=C2)OCCN(C)C)/C3=CC=CC=C3
456201	Active	CC(=O)N1CCN(CC1)C2=CC=C(C=C2)OC[C@H]3CO[C@](O3)(CN4C=CN=C4)C5=C(C=C(C=C5)Cl)Cl
68546	Active	COC1=C(C=C2C(=C1)C(=NC(=N2)N3CCN(CC3)C(=O)C4=CC=CO4)N)OC
68539	Active	CN(C)CCCN1C2=CC=CC=C2CCC3=C1C=C(C=C3)Cl
9279	Active	CC(C)[N+](C)(CCOC(=O)C1C2=CC=CC=C2OC3=CC=CC=C13)C(C)C
1794	Active	C1CCCN(CC1)C2=NC(=C(N=C2Cl)C(=O)N=C(N)N)N
24867541	Inactive	C[C@H]1[C@H]([C@@](C[C@@H](O1)O[C@@H]2[C@H]([C@@H]([C@H](O[C@H]2OC3=C4C=C5C=C3OC6=C(C=C(C=C6)[C@H]([C@H](C(=O)N[C@H](C(=O)N[C@H]5C(=O)NC7C8=CC(=C(C=C8)O)C9=C(C=C(C=C9[C@H](NC(=O)[C@H]([C@@H](C1=CC(=C(O4)C=C1)Cl)O)NC7=O)C(=O)O)O)O)CC(=O)N)NC(=O)[C@@H](CC(C)C)NC)O)Cl)CO)O)O)(C)N)O
24867536	Inactive	C1=CC(=C[N+](=C1)[C@H]2C(C([C@@H](O2)COP(=O)([O-])OP(=O)(O)OC[C@@H]3[C@@H]([C@H]([C@@H](O3)N4C=NC5=C4N=CN=C5N)OP(=O)(O)[O-])O)O)O)C(=S)N

24867535	Inactive	CC([C@H]([C@H]1CNC2=C(N1)C(=O)N=C(N2)N)O)O
24867530	Inactive	C1=CC=C(C(=C1)C(=O)N[C@@H](CCC(=O)[O-])C(=O)[O-])C(=O)[O-]
24867525	Inactive	C1C(C(C(C(C1N)OC2C(C(C(C(O2)CO)O)O)N)O[C@@H]3[C@H]([C@H]([C@H](O3)CO)OC4C(C(C(C(O4)CN)O)O)N)O)O)N
24867522	Inactive	CC(=C)[C@@H]1C2[C@@H]3[C@@]4([C@](C1C(=O)O2)(C[C@@H]5[C@]4(O5)C(=O)O3)O)C
24867518	Inactive	C1C=CN(C=C1C(=O)N)C2C(C(C(O2)COP(=O)([O-])OP(=O)([O-])OC[C@@H]3[C@@H]([C@H]([C@@H](O3)N4C=NC5=C4N=CN=C5N)OP(=O)([O-])[O-])O)O)O
24867514	Inactive	CSC1=NC2=C(C(=N1)N)N=CN2[C@H]3[C@@H]([C@H]([C@H](O3)COP(=O)([O-])OP(=O)([O-])[O-])O)O
24867513	Inactive	CN1CCC[C@H]1C2=CN=CC=C2
24867509	Inactive	[Li+]
24867503	Inactive	CCNC(=O)[C@@H]1[C@@H]([C@H]([C@@H](O1)N2C=NC3=C2N=CN=C3N)O)O
24867493	Inactive	C[C@]12CCC3C([C@@H]1CCC2=O)CC=C4[C@@]3(CC[C@@H](C4)OS(=O)(=O)[O-])C
24867492	Inactive	CC[C@@H](C)C1CN[C@@H]([C@H]1CC(=O)O)C(=O)O
24867487	Inactive	CCNC(=O)[C@@H]1[C@@H]([C@H]([C@@H](O1)N2C=NC3=C2N=C(N=C3N)NCCC4=CC=C(C=C4)CCC(=O)O)O)O
24867481	Inactive	CCCC
24867470	Inactive	C1=CN2[C@H]3[C@H]([C@H]([C@H](O3)CO)O)OC2=NC1=N
24867469	Inactive	C[C@]12CCC3C([C@@H]1CC[C@@H]2O)CCC4[C@@]3(CC[C@H](C4)O)C
24867466	Inactive	CNC(=O)[C@@H]1[C@@H]([C@H]([C@@H](O1)N2C=NC3=C2N=CN=C3N)O)O
24867465	Inactive	CC1(O[C@@H]2[C@H](C(O[C@@H]2O1)[C@@H](CO)O)OCCCN(C)C)C
24848913	Inactive	CC(=NCCCC(C(=O)O)N)N
23682212	Inactive	CC(=O)[C@H]1CCC2[C@@]1(CCC3C2CC=C4[C@@]3(CC[C@@H](C4)OS(=O)(=O)[O-])C)C
23681235	Inactive	C1CC(C(C2=CC=CC=C2C1)O)CC(=O)[O-]
23681234	Inactive	C1=CC(=CC=C1C(/C=C/C(=O)[O-])O)Cl
23681233	Inactive	CCCCCN(CCCCC)C(=O)C(CCC(=O)[O-])NC(=O)C1=CC(=C(C=C1)Cl)Cl
23681059	Inactive	C[C@@H](C1=CC2=C(C=C1)C=C(C=C2)OC)C(=O)[O-]
23679632	Inactive	C1=CC=C(C=C1)C2(C(=O)NC(=N2)[O-])C3=CC=CC=C3
23676659	Inactive	CCCCCC(CCCC(=O)[O-])O
23668244	Inactive	COC1=C(C=CC(=C1)C(CO)O)OS(=O)(=O)[O-]
23663954	Inactive	C1=CC(=CN=C1)CC2=CC3=C(C=C2)OC(=C3)C(=O)[O-]
16760703	Inactive	CCCC(CCC)C(=O)[O-]
16759251	Inactive	CC(=O)NCCC(=O)C1=C(C=CC(=C1)OC)CN=O
16757702	Inactive	C[N+]1(C2CC(CC1C3C2O3)OC(=O)C(CO)C4=CC=CC=C4)C
16219752	Inactive	COC(=O)C(CCCN=C(N)N[N+](=O)[O-])N
13830713	Inactive	CNC1=NC=NC2=C1N=CN2[C@H]3[C@@H]([C@H]([C@H](O3)CO)O)O
12997925	Inactive	C1=CC=C(C=C1)NC2=NC=NC3=C2N=CN3[C@H]4[C@@H]([C@H]([C@H](O4)CO)O)O
12906333	Inactive	C[N+]1(C2CC(CC1C3C2O3)OC(=O)C(CO)C4=CC=CC=C4)C
11957723	Inactive	C[C@]12CCC3C(C1CC[C@@H]2NCCCCCCN4C(=O)C=CC4=O)CCC5=C3C=CC(=C5)OC
11957708	Inactive	CN(C)CCCCSC(=N)N
11957705	Inactive	CCN(CC)CCOC(=O)C1=CC(=C(C=C1OC)N)Cl
11957702	Inactive	CC1C2=C(C(=O)N(C1=O)C)N=CN2
11957691	Inactive	C[C@H]1[C@@H]([C@H]([C@H]([C@@H](O1)OP(=O)(NC(CC(C)C)C(=O)NC(CC2=CNC3=CC=CC=C32)C(=O)[O-])[O-])O)O)O
11957681	Inactive	CNC1=CC(=NC(=N1)NC)NS(=O)(=O)C2=CC=C(C=C2)N
11957668	Inactive	CCCN1CCO[C@H]2[C@H]1COC3=C2C=C(C=C3)O
11957663	Inactive	C1C[C@H]([C@H](NC1)C(=O)O)C(=O)O
11957648	Inactive	C[N+](C)(C)CCC(=O)C1=CC=CC2=CC=CC=C21
11957647	Inactive	CC1=CC=C(C=C1)C(=O)O[C@H]([C@@H](C(=O)O)OC(=O)C2=CC=C(C=C2)C)C(=O)O
11957639	Inactive	C1CC(NC(C1)CCN)CCN
11957622	Inactive	C[Se]C[C@@H](C(=O)O)N
11957621	Inactive	COC1=C(C=CC(=C1)CCN)O
11957614	Inactive	CC(=O)O
11957608	Inactive	CC(C)(CCP(=O)(O)O)C(=O)O

11957601	Inactive	CN1C=C(N=C1)CCN
11957593	Inactive	CC1=CNN=C1
11957581	Inactive	CC(=NCCCC[C@H](C(=O)O)N)N
11957565	Inactive	CC(=O)O
11957562	Inactive	CC(C)[N+]1(C2CCC1CC(C2)C(=O)OC(CO)C3=CC=CC=C3)C
11957558	Inactive	CC1=C(C(=O)NO1)CC(C(=O)O)N
11957555	Inactive	C1CNCC(=C1)C(=O)O
11957538	Inactive	CC(C)C[C@@H](C(=O)O)NC(=O)[C@@H]([C@@H](CC1=CC=CC=C1)N)O
11957537	Inactive	CO[C@H]1CC=C2CCN3[C@]2(C1)C4=C(CC3)COC(=O)C4
11957520	Inactive	C1=C2C(=CC(=C1Cl)Cl)N=C(N2)[C@H]3[C@@H]([C@@H]([C@H](O3)CO)O)O
11957517	Inactive	CC(C)(C)NCC(COC1=CC=CC2=C1NC(=O)N2)O
11957491	Inactive	CN1C(=NC(=O)C(=N1)[O-])SCC2=C(N3C([C@@H](C3=O)NC(=O)/C(=N\OC)/C4=CSC(=N4)N)SC2)C(=O)[O-]
11957485	Inactive	C[N+](C)(C)COP(=O)([O-])OP(=O)([O-])OC[C@@H]1[C@H]([C@H]([C@@H](O1)N2C=CC(=NC2=O)N)O)O
11957478	Inactive	C[N+](C)(C)CCOC(=O)CBr
11957464	Inactive	CCN(CC)C1=NC=NC2=C1N=CN2[C@H]3[C@@H]([C@@H]([C@H](O3)COP(=O)([O-])OP(=O)(C(P(=O)(O)[O-])(Br)Br)[O-])O)O
11957441	Inactive	C1CCN(C1)CCOC(=O)CC2=CC(=C(C=C2)Cl)Cl
11953777	Inactive	C1[C@H]([C@@H]([C@H](N1)CO)O)O
11665606	Inactive	COC1=C(C=C(C=C1)CCN)O
11601888	Inactive	C[C@@H]1O[C@@H](CS1)C[N+](C)(C)C
10404739	Inactive	C(CC(=O)N[C@H](CSN=O)C(=O)NCC(=O)O)[C@@H](C(=O)O)N
10018826	Inactive	CCCN1CCCC2C1CC3=CN=C(N=C3C2)N
9964741	Inactive	CN(C)C(=NCCC[C@@H](C(=O)O)N)N
9922558	Inactive	CC(CC1=CNC2=C1C=C(C=C2)O)N
9884487	Inactive	CC[N+](CC)(CC)CC(=O)NC1=C(C=CC=C1C)C
9855833	Inactive	C[C@H](CC1=CC=CC=C1)NCC#C
9836150	Inactive	CC1=C(C2=C(N1)C=CC(=C2)O)CCN
9795082	Inactive	CC1=C(C(=CC=C1)C)NC(=O)C[N+](C)(C)C
9604977	Inactive	CC1=N/C(=N\NC2=CC=C(C=C2)C(=O)[O-])/C(=C(C1=O)C=O)COP(=O)([O-])[O-]
9549280	Inactive	CC(=O)/C(=C(/NC1=C(C=CC(=C1)Br)Br)\O)/C#N
6918215	Inactive	CCCCCCCCCCCCCCCCCOC[C@H](COP(=O)([O-])OCC[N+](C)(C)C)OC
6917797	Inactive	C1CC(CN(C1)CCC=C(C2=CC=CC=C2)C3=CC=CC=C3)C(=O)O
6917794	Inactive	CCCN(CCC)C1CCC2=C(C1)C(=CC=C2)O
6604094	Inactive	C1CN(CC2=CC=CC=C21)C(=N)[NH3+]
6603931	Inactive	CCCN1C2=C(C(=O)N(C1=O)CCC)NC(=N2)C3=CC=C(C=C3)OCC(=O)NC4=CC=C(C=C4)C#N
6603901	Inactive	CCCC1=C(C=CC(=C1O)C(=O)C)OCCCOC2=CC=C(C=C2)OCC(=O)O
6603697	Inactive	C(/C=C\C(=O)O)N
6532796	Inactive	CC1=N/C(=N/NC2=C(C=C(C=C2)S(=O)(=O)[O-])S(=O)(=O)[O-])/C(=C(C1=O)C=O)COP(=O)([O-])[O-]
6440459	Inactive	C1COCCN1C(=O)NCCNCC(COC2=CC=C(C=C2)O)O
6436473	Inactive	C1CN(CC1)CC#CCN2C(=O)CCC2
6419997	Inactive	C1=C(NC=N1)CCNC(=O)CCN
6419304	Inactive	COC1=CC=C(C=C1)C2=N[N+](=C(C=C2)N)CCCC(=O)O
6409633	Inactive	C/C(=N\O)/C(=O)C
6324610	Inactive	CCOC(=O)C12CC1/C(=N/O)/C3=CC=CC=C3O2
6093162	Inactive	CC1C(O1)P(=O)([O-])[O-]
6093160	Inactive	C1=CC(=CC(=C1)NC(=O)NC2=CC=CC(=C2)C(=O)NC3=C4C(=CC(=CC4=C(C=C3)S(=O)(=O)[O-])S(=O)(=O)[O-])S(=O)(=O)[O-])C(=O)NC5=C6C(=CC(=CC6=C(C=C5)S(=O)(=O)[O-])S(=O)(=O)[O-])S(=O)(=O)[O-]
5702253	Inactive	C1CNCC2=C1C(=O)NO2
5702251	Inactive	C1=CC(=CC=C1C(=O)NCCN)Cl
5702250	Inactive	CCCCCCCCCCCCCCCCCC(=O)OC(CC(=O)O)C[N+](C)(C)C

5702214	Inactive	CCN(CC)CCNC(=O)C1=C(C=CC(=C1)S(=O)(=O)C)OC
5702206	Inactive	CC(=O)O
5702160	Inactive	C1=C(N=C(S1)N=C(N)N)CSCC/C(=N/S(=O)(=O)N)/N
5462653	Inactive	C(CN)C#N
5353800	Inactive	CN(C)C(=O)/N=N/C(=O)N(C)C
5353788	Inactive	CCO/C(=N/C1=C[N+](=NO1)N2CCOCC2)/[O-]
5312115	Inactive	C[C@H](CC1=CC2=C(C=C1)OC(O2)(C(=O)[O-])C(=O)[O-])NC[C@@H](C3=CC(=CC=C3)Cl)O
5311302	Inactive	CNC1=NC=NC2=C1N=CN2[C@H]3C[C@@H]([C@H](O3)COP(=O)(O)[O-])OP(=O)(O)[O-]
5310987	Inactive	C(/C=C/C(=O)O)N
5310956	Inactive	C1[C@@H]([C@H]1C(=O)O)[C@@H](C(=O)O)N
5284443	Inactive	CNC[C@@H](C1=CC(=CC=C1)O)O
5282759	Inactive	CCCCCCC/C=C\CCCCCCCCC(=O)O
5281708	Inactive	C1=CC(=CC=C1C2=COC3=C(C2=O)C=CC(=C3)O)O
5281670	Inactive	C1=CC(=C(C=C1O)O)C2=C(C(=O)C3=C(C=C(C=C3O2)O)O)O
4549312	Inactive	CC1=CC=CC=C1CNC2=NC=NC3=C2N=CN3C4C(C(C(O4)CO)O)O
4234241	Inactive	C1CC2C(C1)C3CC2CC3OC(=S)[S-]
4177957	Inactive	CC1=C(C2=C(N1C(=O)C3=CC=CC=C3)C=CC(=C2)OC)CC(=O)O
3870203	Inactive	C1=C(C=C(C(=C1[N+](=O)[O-])O)O)[N+](=O)[O-]
3864541	Inactive	CCN1C=C(C(=O)C2=C1N=C(C=C2)C)C(=O)[O-]
3519541	Inactive	CC(C)(C)C1=CC(=CC(=C1O)C(C)(C)C)CC(C)(C)C=O
3074827	Inactive	CC1(CCC(CC1)NC(=O)[C@H](CCC(=O)O)N)C
3040551	Inactive	C[C@H]([C@H](C1=CC=C(C=C1)O)O)NCCC2=CC=C(C=C2)O
3035523	Inactive	C1=CC(=CC=C1CN2C=CNC2=S)O
3033332	Inactive	CN/C(=C\[N+](=O)[O-])/NCCSCC1=CC=C(O1)CN(C)C
2837663	Inactive	CC(C)(C(COC1=CC=CC2=C1N(C(=O)C=C2OC)C)O)O
2794990	Inactive	C(CCN=C(N)N)CN
2735510	Inactive	C[N+](C)(C)CC=O
2734952	Inactive	C(=NN)(N)N
2734687	Inactive	C(=NN)(N)N
2733517	Inactive	C1[C@@H](N[C@@H]1C(=O)O)C(=O)O
2733277	Inactive	COC(=O)C(CC1=CC=C(C=C1)Cl)N
2724466	Inactive	C1CCC(CC1)(C(=O)O)N
2723891	Inactive	C(CC(=O)O)[C@@H](C(=O)O)N
2723890	Inactive	COC1=C(C=CC(=C1)C(CO)O)O
1617430	Inactive	C(CC[C@H](C(=O)O)N)CCP(=O)(O)O
1549098	Inactive	C([C@@H](C(=O)O)N)S(=O)O
736715	Inactive	C1=C(NC=N1)/C=C/C(=O)O
689043	Inactive	C1=CC(=C(C=C1/C=C/C(=O)O)O)O
688272	Inactive	CCN1CCC[C@H]1CNC(=O)C2=C(C=CC(=C2)S(=O)(=O)N)OC
657346	Inactive	C[C@H]1OC[C@H](O1)C[N+](C)(C)C
656765	Inactive	CCN1C2=CC(=C(C=C2NC1=O)Cl)Cl
656717	Inactive	C[C@]1(CC2=CC(=C(C(=C2C1=O)Cl)Cl)OCC(=O)O)C3CCCC3
451515	Inactive	CC1=CN(C(=O)NC1=O)[C@H]2C[C@H]([C@H](O2)CO)N=[N+]=[N-]
449215	Inactive	C1[C@@H](C(=O)NO1)N
447196	Inactive	C1=C(C(=O)NC(=O)N1C[C@@H](C(=O)O)N)I
446727	Inactive	C1[C@@H]([C@H](O[C@H]1N2C=C(C(=O)NC2=O)/C=C/Br)CO)O
443586	Inactive	C1=C(C=C(C=C1O)O)[C@@H](C(=O)O)N
443239	Inactive	[C@H]([C@@H](C(=O)O)O)(C(=O)O)N
442897	Inactive	CC(CO)(C1CC2=C(C3=CC=CC=C3N=C2O1)OC)O
441350	Inactive	C[C@](CC1=CC=C(C=C1)O)(C(=O)O)N
441334	Inactive	CC(C)(C)NCC(C1=CC(=CC(=C1)O)O)O

441333	Inactive	CC(C)NCC(C1=CC(=CC(=C1)O)O)O
440005	Inactive	C(C[C@@H](C(=O)O)N)CN=C(N)N[N+](=O)[O-]
439744	Inactive	C1=CC(=C(C=C1C[C@@H](C(=O)O)N)I)O
439280	Inactive	C1=CC2=C(C=C1O)C(=CN2)C[C@@H](C(=O)O)N
433294	Inactive	[Li+]
377339	Inactive	C1C(C(C(C(N1)CO)O)O)O
260390	Inactive	CNCCC1=CNC2=C1C=C(C=C2)O
205536	Inactive	C1=C(ONC1=O)CN
198382	Inactive	C1C(C=CC=C1C(=O)O)N
194216	Inactive	CC1=C(N2[C@@H]([C@@H](C2=O)NC(=O)C(C3=CC=CC=C3)N)SC1)C(=O)O
175540	Inactive	CC(C)NC[C@@H](COC1=CC=C(C=C1)CC(=O)N)O
169743	Inactive	CCN1CCC2=C(CC1)OC(=N2)N
169373	Inactive	CC1=C(N2[C@@H]([C@@H](C2=O)NC(=O)C(C3=CCC=CC3)N)SC1)C(=O)O
167529	Inactive	C=CC[C@@H](C(=O)O)N
160453	Inactive	C1=CC=C2C(=C1)C(=O)N(C2=O)[C@@H](CCC(=O)O)C(=O)O
160436	Inactive	C1=CC2=C(C=C1O)C(=CN2)CCN
157991	Inactive	CC(=O)N[C@H](C(=O)O)C(C)(C)SN=O
155107	Inactive	C1CNCC=C1C(=O)O
145685	Inactive	C1=C(NC=N1)CC(=O)O
135313	Inactive	C(CC(=O)O)[C@@H](C(=O)O)N=C(N)N
120729	Inactive	CC1(CCCC(N1C)(C)C)C
114924	Inactive	CCCCCCCCCCCCCCSCC(=O)O
107812	Inactive	C(CS(=O)O)N
104766	Inactive	C1C[C@](C[C@@H]1C(=O)O)(C(=O)O)N
104762	Inactive	C1=NC(=C(N1[C@H]2[C@@H]([C@@H]([C@H](O2)CO)O)O)O)C(=O)N
102542	Inactive	COC1=C(C=CC(=C1)C(CN)O)O
102484	Inactive	C1=CC(=CC=C1C(CN)O)O
99562	Inactive	CN1C=NC2=C1C(=O)N(C(=O)N2C)CC#C
97587	Inactive	C(CN)CP(=O)(O)O
92913	Inactive	CC(CC1=CC=CC=C1)N(C)CC#C
92222	Inactive	C1=CC(=C(C=C1C[C@H](C(=O)O)N)O)O
92136	Inactive	C(C[C@@H](C(=O)O)N)CC(=O)O
89034	Inactive	CS(=N)(=O)CC[C@@H](C(=O)O)N
84003	Inactive	C1CN2C(=CC=C2C(=O)C3=CC=CC=C3)C1C(=O)O
80289	Inactive	C1=CC=C(C=C1)C(=N)N
74724	Inactive	C[N+](C)(C)CCO
71417	Inactive	CCN(CC)CCNC(=O)C1=CC=C(C=C1)NC(=O)C
69398	Inactive	C(CCN)CC(=O)O
66449	Inactive	C1=CC(=CC=C1CCN)O
66091	Inactive	C1=C(NC=N1)C[C@@H](C(=O)O)N
66068	Inactive	CCN(CC)CCNC(=O)C1=CC=C(C=C1)N
57004	Inactive	C(CC(C(F)F)(C(=O)O)N)CN
55918	Inactive	C1CC(=O)NN=C1C2=CC=C(C=C2)N3C=CN=C3
40958	Inactive	C1C(=C(N2[C@H](S1)[C@@H](C2=O)NC(=O)C(C3=CC=CC=C3)N)C(=O)O)Cl
40632	Inactive	CC1=CN(C(=O)C=C1)C2=CC=CC=C2
40539	Inactive	C([C@@H](C(=O)O)N)N1C(=O)NC(=O)O1
39912	Inactive	C[C@@H](C1=CC=C(C=C1)CC(C)C)C(=O)O
39859	Inactive	CC(C)(C)NCC(C1=CC(=C(C=C1)O)CO)O
39562	Inactive	C1=CC=C2C(=C1)C(=NN2CC3=C(C=C(C=C3)Cl)Cl)C(=O)O
39214	Inactive	CC1=NC=C(C(=N1)N)CNC(=O)N(CCCl)N=O
31307	Inactive	C[C@]12C[C@@H]([C@]3([C@H]([C@@H]1C[C@H]([C@@]2(C(=O)CO)O)O)CCC4=CC(=O)C

		=C[C@@]43C)F)O
31195	Inactive	C(P(=O)(O)[O-])(P(=O)(O)[O-])(Cl)Cl
24066	Inactive	C1C[C@@H](O[C@@H]1CO)N2C=CC(=NC2=O)N
22880	Inactive	CN[C@H](CC(=O)O)C(=O)O
22475	Inactive	C[N+](C)(C)CCOC(=O)CCC(=O)OCC[N+](C)(C)C
22411	Inactive	CC(=O)SCC[N+](C)(C)C
18343	Inactive	C[C@@H]1[C@H]([C@H]([C@@H](O1)N2C=C(C(=O)NC2=O)F)O)O
18283	Inactive	CC1=CN(C(=O)NC1=O)[C@H]2C=C[C@H](O2)CO
17882	Inactive	C[N+]1(C2CCC1CC(C2)OC(=O)C(CO)C3=CC=CC=C3)C
16817	Inactive	C[C@H]1[C@@H](C[C@H](O1)C[N+](C)(C)C)O
16486	Inactive	C1CN[C@@H]1C(=O)O
13347	Inactive	CSC(=N)N
12035	Inactive	CC(=O)N[C@@H](CS)C(=O)O
11545	Inactive	C[N+](C)(C)CC(=O)O
11236	Inactive	C(=O)(N)NN
10729	Inactive	C1=NC2=C(NC1=O)NC(=NC2=O)N
10255	Inactive	CC(=C)[C@H]1CN[C@@H]([C@H]1CC(=O)O)C(=O)O
9539	Inactive	C(CCNCCCN)CN
9532	Inactive	C(CCN)CN
9444	Inactive	C1=NC(=NC(=O)N1[C@H]2[C@@H]([C@@H]([C@H](O2)CO)O)O)N
9433	Inactive	CN1C2=C(C(=O)N(C1=O)C)NC=N2
9367	Inactive	CC(C)NNC(=O)C1=CC=NC=C1
9363	Inactive	CCN(CC)C(=O)C1=CC(=C(C=C1)O)OC
9082	Inactive	C(CS)N
8743	Inactive	C1=C(NC(=CC1=O)C(=O)O)C(=O)O
8246	Inactive	CN(C)C(=O)OC1=CC=CC(=C1)[N+](C)(C)C
7550	Inactive	C[N+]1=CC=CC(=C1)OC(=O)N(C)C
7172	Inactive	CNCC(C1=CC=C(C=C1)O)O
6322	Inactive	C(C[C@@H](C(=O)O)N)CN=C(N)N
6252	Inactive	C1=CN(C(=O)N=C1N)[C@H]2[C@H]([C@@H]([C@H](O2)CO)O)O
6207	Inactive	C(COCCOCCN(CC(=O)O)CC(=O)O)N(CC(=O)O)CC(=O)O
6198	Inactive	COC1=CC2=C(C=C1)NC=C2CCN
6172	Inactive	CC[N+](CC)(CC)CCOC1=C(C(=CC=C1)OCC[N+](CC)(CC)CC)OCC[N+](CC)(CC)CC
6114	Inactive	CC(C[N+](C)(C)C)OC(=O)C
6112	Inactive	C1CSS[C@@H]1CCCCC(=O)O
6076	Inactive	C1[C@@H]2[C@H]([C@H]([C@@H](O2)N3C=NC4=C3N=CN=C4N)O)OP(=O)(O1)O
6035	Inactive	C1[C@@H]([C@H](O[C@H]1N2C=C(C(=O)NC2=O)Br)CO)O
5961	Inactive	C(CC(=O)N)[C@@H](C(=O)O)N
5960	Inactive	C([C@@H](C(=O)O)N)C(=O)O
5938	Inactive	C[N+](C)(C)CCCCCC[N+](C)(C)C
5917	Inactive	C1CCC2=NN=NN2CC1
5860	Inactive	C[N+]1(C2CCC1CC(C2)OC(=O)C(CO)C3=CC=CC=C3)C
5849	Inactive	C[N+]1(CCCC1)CCCCC[N+]2(CCCC2)C
5831	Inactive	C[N+](C)(C)CCOC(=O)N
5818	Inactive	C1=C(NC=N1)CCN
5723	Inactive	COC1=C(C=CC(=C1)C2=NNC(=O)C=C2)OC(F)F
5665	Inactive	C=CC(CCC(=O)O)N
5593	Inactive	CCN(CC1=CC=NC=C1)C(=O)C(CO)C2=CC=CC=C2
5520	Inactive	CP(=O)(C1=CC[NH2+]CC1)[O-]
5503	Inactive	CC1=CC=C(C=C1)S(=O)(=O)NC(=O)NN2CCCCCC2
5429	Inactive	CN1C=NC2=C1C(=O)NC(=O)N2C

5426	Inactive	C1CC(=O)NC(=O)C1N2C(=O)C3=CC=CC=C3C2=O
5355	Inactive	CCN1CCCC1CNC(=O)C2=C(C=CC(=C2)S(=O)(=O)N)OC
5335	Inactive	C1=CC=C(C=C1)N2C(=CC=N2)NS(=O)(=O)C3=CC=C(C=C3)N
5242	Inactive	C(=O)(C(=O)[O-])N
5123	Inactive	C1=CN(C(=O)NC1=O)CC(C(=O)O)N
4971	Inactive	CC1=C(C2=CC3=C(C(=C(N3)C=C4C(=C(C(=N4)C=C5C(=C(C(=N5)C=C1N2)C)CCC(=O)O)CCC(=O)O)C)C=C)C)C=C
4943	Inactive	CC(C)C1=C(C(=CC=C1)C(C)C)O
4922	Inactive	CCCN(CCC)C(=O)C(CCC(=O)O)NC(=O)C1=CC=CC=C1
4909	Inactive	CCC1(C(=O)NCNC1=O)C2=CC=CC=C2
4843	Inactive	C1CC(=O)N(C1)CC(=O)N
4838	Inactive	C1CNCCC1S(=O)(=O)O
4779	Inactive	C1=CC=C(C(=C1)CP(=O)(O)O)C2=CC(=CC=C2)CC(C(=O)O)N
4740	Inactive	CC(=O)CCCCN1C(=O)C2=C(N=CN2C)N(C1=O)C
4652	Inactive	C1=CC(=CC=C1CC(C(=O)O)N)Cl
4650	Inactive	C1=CC(=O)C=CC1=O
4488	Inactive	C1=CC(=CC(=C1)NC2=C(C=CC=N2)C(=O)O)C(F)(F)F
4389	Inactive	C1CC(N(C1)C(=O)CCC(=O)O)C(=O)O
4386	Inactive	C1=CC=C(C=C1)NC2=CC=CC=C2C(=O)O
4362	Inactive	CCN1C(=O)C=CC1=O
4353	Inactive	CC(=O)NBr
4201	Inactive	C1CCN(CC1)C2=NC(=N)N(C(=C2)N)O
4197	Inactive	CC1=C(C=C(C(=O)N1)C#N)C2=CC=NC=C2
4038	Inactive	CC1=C(C(=C(C=C1)Cl)NC2=CC=CC=C2C(=O)[O-])Cl
3857	Inactive	C(C(C(=O)O)N)P(=O)(O)O
3845	Inactive	C1=CC=C2C(=C1)C(=O)C=C(N2)C(=O)O
3825	Inactive	CC(C1=CC=CC(=C1)C(=O)C2=CC=CC=C2)C(=O)O
3758	Inactive	CC(C)CN1C2=C(C(=O)N(C1=O)C)NC=N2
3727	Inactive	C(C(=O)N)I
3657	Inactive	C(=O)(N)NO
3454	Inactive	C1=NC2=C(N1COC(CO)CO)NC(=NC2=O)N
3446	Inactive	C1CCC(CC1)(CC(=O)O)CN
3433	Inactive	CC1=NC2=C(N1)C(=O)N(C(=O)N2CC3=CC=CO3)C
3373	Inactive	CCOC(=O)C1=C2CN(C(=O)C3=C(N2C=N1)C=CC(=C3)F)C
3331	Inactive	C1=CC=C(C=C1)C(COC(=O)N)COC(=O)N
3291	Inactive	CCC1(CC(=O)NC1=O)C
3132	Inactive	C1=CC=C(C=C1)CC(CS)C(=O)NCC(=O)O
3125	Inactive	CC(CC1=CC=C(C=C1)O)(C(=O)O)N
3122	Inactive	C(CCC(C(=O)O)N)CCP(=O)(O)O
3019	Inactive	CC1=NS(=O)(=O)C2=C(N1)C=CC(=C2)Cl
2944	Inactive	C1=C(NC(=NC1=O)N)N
2935	Inactive	C(CC(=O)NCS(=O)(=O)O)C(C(=O)O)N
2910	Inactive	C1C2CC(C1C=C2)C3NC4=CC(=C(C=C4S(=O)(=O)N3)S(=O)(=O)N)Cl
2907	Inactive	C1CNP(=O)(OC1)N(CCCl)CCCl
2796	Inactive	CCOC(=O)C(C)(C)OC1=CC=C(C=C1)Cl
2763	Inactive	CC(C)(C(=O)O)OC1=CC=C(C=C1)C2CC2(Cl)Cl
2733	Inactive	C1=CC2=C(C=C1Cl)NC(=O)O2
2727	Inactive	CCCNC(=O)NS(=O)(=O)C1=CC=C(C=C1)Cl
2576	Inactive	CCCC(C)(COC(=O)N)COC(=O)NC(C)C
2519	Inactive	CN1C=NC2=C1C(=O)N(C(=O)N2C)C
2471	Inactive	CCCCNC1=C(C(=CC(=C1)C(=O)O)S(=O)(=O)N)OC2=CC=CC=C2

2331	Inactive	C1=CC=C(C=C1)C(=O)N
2284	Inactive	C1=CC(=CC=C1C(CC(=O)O)CN)Cl
2266	Inactive	C(CCCC(=O)O)CCCC(=O)O
2249	Inactive	CC(C)NCC(COC1=CC=C(C=C1)CC(=O)N)O
2244	Inactive	CC(=O)OC1=CC=CC=C1C(=O)O
2207	Inactive	C(CP(=O)(O)O)C(C(=O)O)N
2196	Inactive	COC1=CC=C(C=C1)C(=O)N2CCCC2=O
2145	Inactive	CCC1(CCC(=O)NC1=O)C2=CC=C(C=C2)N
2141	Inactive	C(CN)CNCCSP(=O)(O)O
2123	Inactive	CN(C)C1=NC(=NC(=N1)N(C)C)N(C)C
2094	Inactive	C1=C2C(=NC=NC2=O)NN1
2083	Inactive	CC(C)(C)NCC(C1=CC(=C(C=C1)O)CO)O
2071	Inactive	C1CC(C2=C1C=C(C=C2)C(=O)O)(C(=O)O)N
1989	Inactive	CC(=O)C1=CC=C(C=C1)S(=O)(=O)NC(=O)NC2CCCCC2
1986	Inactive	CC(=O)NC1=NN=C(S1)S(=O)(=O)N
1893	Inactive	C1=CC2=C(C(=C1)[N+](=O)[O-])NN=C2
1779	Inactive	C1=C(C=C2C(=C1Cl)C(=O)C=C(N2)C(=O)O)Cl
1775	Inactive	C1=CC=C(C=C1)C2(C(=O)NC(=O)N2)C3=CC=CC=C3
1774	Inactive	CC1(CCC=[N+]1[O-])C
1742	Inactive	C1=CC(=CC=C1C(=O)NN)O
1738	Inactive	COC1=C(C=CC(=C1)CC(=O)O)O
1727	Inactive	C1=CN=CC=C1N
1678	Inactive	C(C[N+](=O)[O-])C(=O)O
1676	Inactive	CCCN1C2=C(C(=O)NC1=O)NC=N2
1645	Inactive	C1=CC(=CC(=C1)N)C(=O)N
1641	Inactive	C1=CC(=CC=C1C(CN)CP(=O)(O)O)Cl
1564	Inactive	C1=CC(=CC=C1C(CN)(CS(=O)(=O)O)O)Cl
1390	Inactive	CN1C=CN=C1
1365	Inactive	CN1C2=C(N=C1C3=CC=C(C=C3)S(=O)(=O)O)N(C(=O)N(C2=O)CC=C)C
1340	Inactive	C1=CC2=C(C=CNC2=O)C(=C1)O
1256	Inactive	CC(=C)C1CCC(=CC1)C(=O)O
1245	Inactive	COC1=C(C=CC(=C1)C(C(=O)O)O)O
1233	Inactive	C1=C(ONC1=O)C(C(=O)O)N
1232	Inactive	C1CN(C(=O)C1N)O
1228	Inactive	C1CN(CC(N1)C(=O)O)CCCP(=O)(O)O
1222	Inactive	CC(C1=CC=C(C=C1)C(=O)O)(C(=O)O)N
1216	Inactive	C(CC(C(=O)O)N)CP(=O)(O)O
1123	Inactive	C(CS(=O)(=O)O)N
1066	Inactive	C1=CC(=C(N=C1)C(=O)O)C(=O)O
1046	Inactive	C1=CN=C(C=N1)C(=O)N
903	Inactive	CC(=O)NCCC1=CNC2=C1C=C(C=C2)O
896	Inactive	CC(=O)NCCC1=CNC2=C1C=C(C=C2)OC
650	Inactive	CC(=O)C(=O)C
564	Inactive	C(CCC(=O)O)CCN
401	Inactive	C1C(C(=O)NO1)N
275	Inactive	C(CON=C(N)N)C(C(=O)O)N
178	Inactive	CC(=O)N
119	Inactive	C(CC(=O)O)CN