universität wien

DISSERTATION

Titel der Dissertation

„Similarity based classification studies
for prediction of ABCB1 (P-glycoprotein)
substrates and non-substrates"

Verfasserin

Mag. pharm. Rita Schwaha

angestrebter akademischer Grad

Doktorin der Naturwissenschaft (Dr. rer. nat.)

Wien, im Jahr 2013

Studienkennzahl lt. Studienblatt:	A 091 449
Dissertationsgebiet lt. Studienblatt:	Dr.-Studium der Naturwissenschaften Pharmazie
Betreuer:	Univ.-Prof. Mag. pharm. Dr. rer. nat. Gerhard F. Ecker

Acknowledgements

This work has taken a long time from start to finish and would have been neither started nor finished without the help of a great number of people. First of all great thanks to my supervisor Univ.-Prof. Dr. Gerhard Ecker for giving me the chance of doing a PhD thesis in his research group. He always offered me his support, his experience and his ideas in the course of this thesis. He made it possible for me to combine part-time work in the pharmacy and my research at the university.

Of course life as a PhD student would have been much harder without a lively, intelligent and funny group of co-workers at the university. Especially Lars Richter and Michael Demel have been my companions from the start. Lars was always there with a funny joke and a cheering comment as well as serious scientific advice. Michael presented his insights and jointly we explored ideas for various future strategies. A great "thank you" also to René Weissensteiner for many shared cigarette breaks and for giving me the final push to physically write my thesis at last. Thanks also to the other members of the Pharmacoinformatic Research group who contributed support and social distraction.

My family naturally deserves the biggest praise for supporting me during the whole time. My father was always there with his experiences and his own scientific background. My mother gave me the opportunity to do my thesis in spite of the amount of work that accompanies the founding of a new pharmacy and put my spirits up when needed. Also my brother Stefan and my sister Claudia provided support when nerves failed and always lent an ear when problems occurred. Many thanks also are due to my friends who had to spend an evening once in a while listening to the problems of a PhD student. Also my colleagues at the pharmacy sometimes had to bear the brunt of the stressful life of a part-time PhD student and part-time pharmacist.

Last but not least I acknowledge the financial support of the Austrian Science Fund (FWF).

Ich habe mich bemüht, sämtliche Inhaber der Bildrechte ausfindig zu machen und ihre Zustimmung zur Verwendung der Bilder in dieser Arbeit eingeholt. Sollte dennoch eine Urheberrechtsverletzung bekannt werden, ersuche ich um Meldung bei mir.

Contents

List of Figures

List of Tables

1

Literature Survey - State of the Art

1.1 Physiological function of ABCB1 (P-glycoprotein)

The so-called multidrug resistance (MDR) ATP-binding cassette (ABC) protein family plays an important role in absorption, excretion of drugs and therapeutic drug resistance. The family consists of 48 different transporters and seven subfamilies (A to G) with the most prominent member of this family being ABCB1 (P-glycoprotein, Pgp). It first came to notice as principal cause for multidrug resistance especially considering cancer drugs. Later on other members of the family were revealed as additional major influences thereon, i.e. the multidrug resistance proteins (MRPs, ABCC subtypes) 1-5 and the breast cancer resistance protein (BCRP, ABCG2).[1,2] In this work the focus lies on ABCB1.

ABCB1 is a 170 kDa efflux transporter encoded by the MDR1 and MDR2 genes in humans and its main function is the export of xenobiotic substances. The human MDR1 gene confers the multidrug resistance whereas the MDR2 encoded P-glycoproteins are involved in phosphatidyl choline excretion into bile across the liver. Further on in this work the term ABCB1 is meant for the MDR1 gene product expressed in the polarised epithelial cells at the apical

membrane and localised in the intestine, kidney, Blood Brain Barrier (BBB), blood cerebrospinal fluid barrier (BCSF), blood-testis barrier, pancreas and peripheral immune cells, the adrenal gland and placental trophoblasts.[3]

1.1.1 Intestinal ABCB1 – Effects on oral bioavailability

Oral drug administration is highly preferred as it is cheap, relatively safe and meets patients' needs in terms of compliance. Formerly pharmaceutical industry concentrated on Phase I and Phase II metabolism of drugs whereas nowadays two additional steps, called Phase 0 and Phase III have emerged. Phase 0 involves the modulation of the cellular entry whereas Phase III describes the cellular exit of the detoxified compounds.[4] In these phases ABCB1 among others plays an important part as it is situated in the gut wall mucosa and forms a so-called drug-efflux metabolic alliance with CYP3A4. Herein ABCB1 functions as efflux pump in order to provide CYP3A4 with further access to the available drugs for metabolism in a circle of absorption and efflux. As both are induced by the orphan nuclear Pregnane X receptor (PXR)[5] receptor co-regulation of these two metabolic entities may be a valid hypothesis. Another way altogether may be that other genetic or environmental factors may play an important role. ABCB1 inhibitors selective in this instance could provide better oral bioavailability as for example paclitaxel and HIV protease inhibitors could be administered orally.[6] However, this would bear another danger because potentially narrower therapeutic margins of orally better available drugs[7] are the consequence.

1.1.2 ABCB1 at the Blood Brain Barrier

The blood brain barrier (BBB) represents a physical and metabolic barrier. It is composed among others of a monolayer of brain capillary endothelial cells and tight junctions between those cells. Circulating molecules are presented with two ways of entry: First, lipid-mediated transport by passive diffusion and second catalysed transport. Various membrane transporters are responsible for efflux and influx of essential substances such as amino acids, electrolytes, nucleosides and glucose. In former days drug transport

was thought to function entirely based on physicochemical parameters such as molecular weight, lipophilicity and ionic state whereas nowadays one of the most important functions of ABCB1 comprises the protection of the brain.[8] However, it has to be borne in mind that the higher the affinity of the compound to the transporter the higher the extrusion rate. Weak affinity substrates to ABCB1 in competition with high affinity substrates may allow the weak affinity substrates entry to the brain which presents a scenario for drug-drug-interactions. Inhibited ABCB1 function increases CNS exposure to a drug 10-100 fold. Digoxin, a known ABCB1 substrate, combined with the ABCB1 modulator quinidine decreased renal and intestinal clearance of digoxin and increased plasma levels, in a way that can be fatal for a drug of this narrow therapeutic margin.[9]

1.1.3 Other physiological functions of ABCB1

ABCB1 is also placed in the placenta, the functional barrier between mother and foetus. It is expressed in the apical membrane on the maternal side thereby protecting the foetal blood circulation. In most cases low to none foetal blood concentration of drugs would be ideal but especially when regarding HIV shortly before birth high foetal blood levels would be preferable.[2] Other ABCB1 functions include the hepatobiliary pathway as it is found in the canalicular membrane of hepatocytes where it extrudes substances from the blood into the bile towards faecal excretion.[10] In the kidney ABCB1 function is not fully clear. An excretory function in the proximal tubes would fit into the picture although controversial results give rise to further speculation. Another function of ABCB1 altogether may be in tissue reparation as studies with stem cells show.[11]

1.1.4 ABCB1 and diseases

In the field of cancer therapy, epilepsy, depression and HIV the effect of drugs in the brain and also in tumours is highly desirable and especially in epilepsy and cancer therapy often meets with drug resistance as most of the therapeutic compounds are substrates of ABCB1. Interestingly a marked

induction of ABCB1 expression and as follows further extrusion of drugs can be observed in these diseases. It remains to be seen whether the induction of ABCB1 expression is due to the diseases (intrinsic) or due to the exposure to the drugs (acquired).[5,8,12] The reverse effect can be seen in Alzheimer Disease (AD) patients as ABCB1 also extrudes β-amyloid peptide (Aβ) from the brain. Aβplaques accumulated in the extracellular compartment of the brain parenchyma are significant for AD. In AD patients a decrease in ABCB1 expression can be observed. In this case better understanding of signalling pathways and regulation could open a window for AD treatment.[13]

As mentioned above ABCB1 also plays a role concerning HIV therapy. ABCB1 besides extrusion of drugs from the brain is also expressed in the CD4+ which are the major target for HIV. Therapeutic success is dependent on exposure of the virus to the therapeutic agent and therefore intracellular drug concentration could bring better responses to therapeutic effort than drug concentration in the plasma.[6]

1.1.5 ABCB1 and natural products

Induction and decrease of ABCB1 expression can also be observed with St. John Wort, a natural product used in treatment of mild depression and agitation[5] and cannot be eliminated for food or environmental sources. Emphasis has been put on the fact that co-medication based on plants or extracts may have its impact on ABCB1.[14] Mistletoe and Carica papaya have been shown to act as ABCB1 inhibitors and therefore also herbal co-medication has to be monitored regarding ABCB1 interaction potential.

1.1.6 Substrates of ABCB1

Substrates of ABCB1 differ widely as ABCB1 is a very promiscuous transporter, which is of course necessary for its functions but extremely difficult when seen as an anti-target in drug development. Overall it can be said that substrates of ABCB1 often are hydrophobic or amphiphilic molecules and often also cationic.[15] Their size is ranging from 200 Da to almost 1900 Da. These substrates include a wide variety of chemotherapeutic agents of natural

origin such as anthracyclines (doxorubicin), vinca alkaloids, epidophyllotoxins and taxanes[3] as well as antidepressants, antivirals, antibiotics, hormones etc. Flavonoids for example interact with ABC proteins within their nucleotide binding-domain or when hydrophobic within the transmembrane domains. Stilbenes on the other hand probably interact directly with substrate binding sites and phenothiazines act as inhibitors of ABCB1 and at the same time as stimulators of ABCC1 (MRP).[16] The extreme diversity of transported substances represents a real challenge to medicinal chemists as in one instance some drugs should not be able to enter the CNS but nevertheless be orally available therefore should be designed as weak ABCB1 substrates. However, in other cases, i.e. epilepsy, HIV, depression, ... compounds should be able to cross the BBB therefore should be designed ABCB1 non-substrates.

ABCB1 substrates are hard to characterise as their number is so varied but common properties seem to be hydrophobicity, a weakly amphipathic nature, not always an aromatic ring and a positively charged N atom. They tend to have planar aromatic domains and tertiary amino groups although the last two requirements do not seem really essential as studies have proven.[4] Due to their mostly hydrophobic nature they passively diffuse into the membrane enabling their uptake. When purposely combining high affinity substrates with low affinity substrates chances are high that partial inhibition or occupation of ABCB1 by the high affinity substrates is enough to allow the low affinity substrates entering by undisturbed passive diffusion.

The interaction between transporter and substrate takes place directly in the plasma membrane and substrates then exit into the cytosol. Two hypotheses exist regarding the precise function of ABCB1. The first hypothesis is the so-called hydrophobic vacuum cleaner model where substrates enter the transporter from the lipid bilayer and consequently are expelled directly into the extracellular space. The other model is the flippase model where substrates are transported from the cytoplasmic to the extracellular membrane leaflet. They are rehydrated and diffuse back into the extracellular fluid.[17] In the process of expulsion ABCB1 undergoes a conformational change. A

two-step model also exists for exporting certain substrates where transfer from the binding site of the ligand to a second site inside of ABCB1 near the exoplasmic side of the membrane and subsequent release[18] takes place.

1.1.7 Inhibitor design

As soon as multidrug resistance became known research for possible ABCB1 inhibitors started. The first inhibitors of generation 1 comprised cyclosporine A and Verapamil though side effects proved too severe combined with their low affinity profile. Second generation inhibitors were developed as analogues though their specificity was not adequate as they also inhibited CYP450 enzymes thereby again causing toxic side effects. Third generation inhibitors like tariquidar seem promising and it remains to be seen how clinical trials fare. They are non-competitive inhibitors and more specific than former generations. Nevertheless inhibition of ABCB1 besides some advantages also bears many risks. Most patients are multi-morbid and take many medications at once thereby enhancing the risk of toxic side effects. Other possibilities for exported drugs contain liposomal entities or nano particles where the drug is smuggled through the BBB in a so-called trojan horse.[12]

Another important aspect in inhibitor design and ABCB1 in general are the gender differences as man and woman generally have different expression rates for ABCB1 in tissues and metabolism. Possible hormonal side effects like regulation of ABCB1 or drug-drug interactions have to be monitored regarding the high number of women taking oral contraceptives.[19]

1.2 Structure of ABCB1

Opposed to Eukarya, Bacteria and Archaea express their ABC transporters as half-transporters with one nucleotide binding domain (NBD) and one transmembrane domain (TMD) on a single polypeptide chain. Two of these chains then come together into a functional homo- or heterodimer. Eukarya on the other hand express their ABC exporters often as a single polypeptide chain containing the essential four domains. All ABC transporters are

composed of two nucleotide binding domains and two transmembrane domains.[20,21] The NBDs are highly conserved throughout the family and include the Walker A motif, Walker B motif and the C signature providing the family with their name. The whole transporter consists of 1280 amino acids with 610 amino acids for each monomer and a linker region of about 60 amino acids. Each monomer consists of 6 helices, in total 12 for the functional transporter.

1.2.1 Nucleotide binding domain

The nucleotide binding domains are responsible for the binding and the hydrolysis of ATP thereby representing the motor of the cell. Their length lies between 200 and 300 amino acids and they show a conserved fold whose arrangement is called "head-to-tail". The key feature of the "head-to-tail" arrangement is the conserved sequence motif at the shared interface of the two NBDs.[22] As a consequence the nucleotide-binding site is positioned at the subunit-subunit interface.[23] A nomenclature of "open" and "closed" NBD conformations is widely used meaning the nucleotide-free and the ATP-bound states of the NBDs.

1.2.2 Transmembrane binding domains

Opposed to the conserved state of the nucleotide binding domains the transmembrane domains though existent throughout the family show far more diversity. They are responsible for the essential function of ABCB1 like substrate binding and substrate export out of the cell. The six helices are connected by three extracellular and 2 intracellular loops.[24] (Figure 1.1).

1.2.3 Crystal structure

Due to enormous problems in producing, purifying and crystallising these membrane proteins crystal structures have been scarce to non-existent for a long time. The only structure available has been retracted in 2006[25] due to software problems but has fortunately been immediately replaced by a new

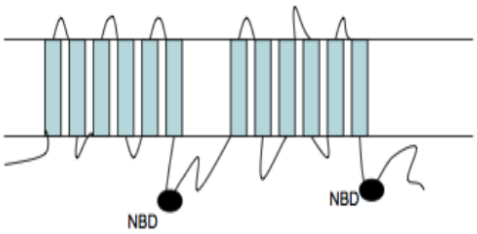


Figure 1.1: Schematic structure of ABCB1

study by Dawson and Locher.[26] They crystallized the structure of a bacterial ABC transporter (Sav1866) in complex with ADP at a high resolution of 3.0 Å. Herein they describe the arrangement in consensus with the widely supposed view of a funnel like appearance for ABCB1.[27] The transmembrane helices yield the picture of two discrete "wings" pointing away from one another towards the cell exterior and therefore represents the outward conformation. The TMDs extend beyond the membrane and reach far into the cytoplasm.[28] Each of these wings consists of two to three helices from each subunit - TM1 and TM2 from subunit one, TM3-TM6 from subunit 2.

As ATP binding and subsequent ATP hydrolysis cause major conformational changes in the NBD these are transmitted to the TMDs through non-covalent interactions at the shared interface. The common interface is mainly comprised of two intracellular loops which bear the bulk of the contacts. In future reference they will be called "coupling helices". A most interesting aspect is the fact that in the Sav1866 structure the coupling helices reach across and contact primarily the nucleotide binding subunit of the opposite subunit. Helix 1 provides an interface to both NBDs but helix 2 only interacts with the opposite subunit.[26]

Andrew Ward and colleagues[29] followed suit and published four different X-ray structures of MsbA with two nucleotide bound products and two crystal structures without nucleotides. The most astonishing part provided the open apo-form presented in this study as it showed an open cavity between the two NBDs of at least 50 Å. This seems a bit heavy to digest as electron microscopy experiments favoured a closer connection of the two NBDs even in a nucleotide free state.[30] Only further studies can clarify this aspect in con-

formational changes.[31] However, the domain swapping as presented by the Sav1866 structure could be supported by these structures with a resolution of about 4.5 to 5 Å.

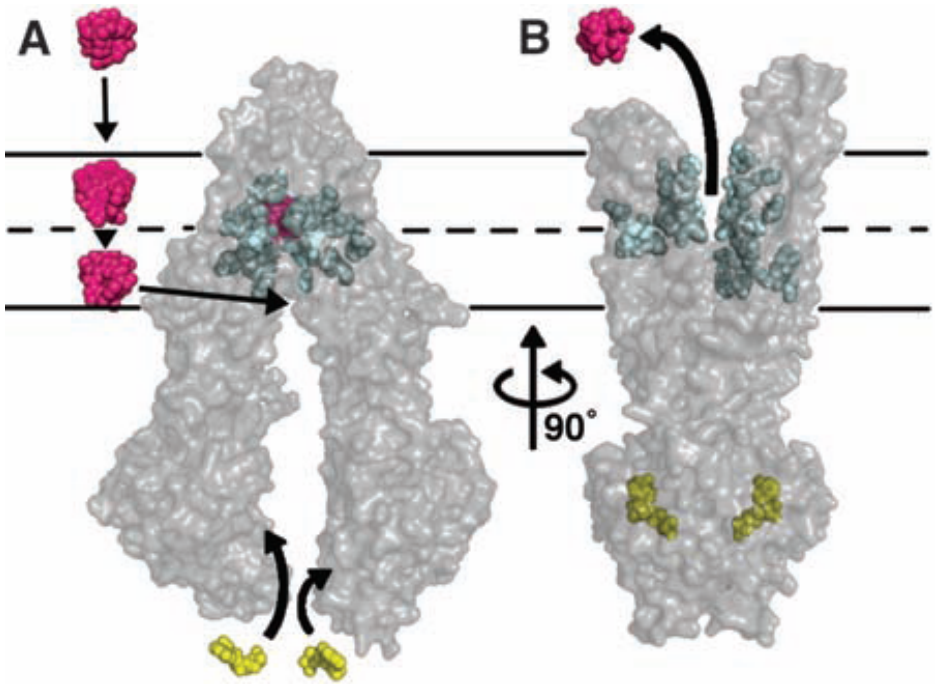


Figure 1.2: Crystal structure of ABCB1; taken from Aller *et al.*[32] Copyright Science

In 2009 the first mammalian X-ray structure of an ABC transporter was published by Aller *et al.*[32] (Figure 1.2). Once more the "wing" - V like structure already observed in Sav1866 could be shown. For the first time also inhibitors were available as bound into the transporter with a resolution of about 4.4 Å. Only recently another crystal structure of ABCB1 with resolution of 3.4 Å, this time from Caenorhabditis elegans, has seen the light and is in general accord with the previous findings.[33] Results of Klepsch *et al.*[24] published in a recent review indicate that the binding of the large inhibitors present in the structure hardly shows any effect at the protein structure. This is puzzling as especially ABCB1 always was considered highly flexible.

1.2.4 Drug binding pocket

The poly-specificity of ABCB1 has been deeply lamented as well as highly praised in recent years. With the published crystallized structures the secret hereof is closer to unravelling. Several TMDs of ABCB1 form a large aqueous cavity closed at the cytoplasmic face of the membrane but open to the extracellular milieu. Also openings in the lipid phase can be observed thereby allowing substrates to enter.[34] The TMDs are arranged similar to lining of a pore and shaped like a funnel and at the cytoplasmic side TM2-TM11 and TM5-TM8 come together.[35] The volume of the internal cavity takes up to 6000 Å^{32} (Figure 1.3) and is therefore large enough to accommodate bulky molecules and also more than one molecule at a time as has been proven before.[36] The drug binding cavity consists mostly of hydrophobic to aromatic residues at the upper side whereas in the lower part of the cavity polar and charged residues are localised.[32] This composition of the cavity is in accord with the famous poly-specificity displayed by ABCB1. After diffusion of hydrophobic substrates into the membrane two portals of ABCB1 in the inward open conformation are available.

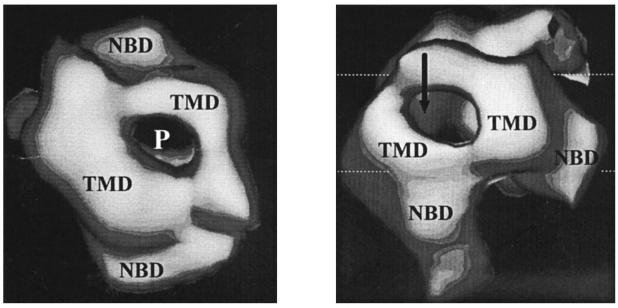


Figure 1.3: Drug binding cavity of ABCB1; taken from Rosenberg *et al.*,[34] Pgp at 2,5 Å. Copyright Journal of Biological Chemistry

These portals are open throughout the width of the transporter and specific substrates are then bound to the transporter. As mentioned previously it has been determined that at least two drugs may bind simultaneously into this large drug binding pocket and there exist at least three interacting

substrate-binding sites, two referred to as the R-site for competitive substrates like rhodamine and the H-site for competitive substrates like Hoechst 33342. Another site altogether is available for modulator binding. These multiple drug binding sites suggest that one substrate binding could promote transport of the other substrates either through cooperative binding or two occupied sites stimulate ATP binding prior to NBD closing conformation.[37] The hypothesis exists that the cavity giving partial access throughout the membrane also hides other binding regions located closer to the NBDs and intracellular domains. Structurally diverse ligands can be accommodated by stacking and cation-π interactions and by hydrophobic and hydrogen bonding interactions.[38] Loo and Bartlett[27] proposed the idea of a "substrate-induced fit" mechanism, meaning essentially that upon substrate binding in order to accommodate the substrate in the binding pocket slight or major conformational changes would take place. They stipulated that common residues could be involved in the binding of different substrates and hence account for the extreme poly-specificity of ABCB1. Although part of this phenomenon may still be true it has to be mentioned that major conformational changes take place in any case as the transporter changes from inward conformation to outward conformation in order to export the substrate.

Sharom[35] highlighted the importance of the lipid bilayer in function and substrate selection of ABCB1. Lipids may interact indirectly or directly with the function of the NBDs through direct interaction with the TMDs. It was found that a direct correlation exists between lipid partitioning ability and drug binding affinity. The higher the partitioning of a drug into the lipid, the higher the apparent binding affinity.

1.2.5 Catalytic Cycle

Formerly ABCB1 was thought to function primarily as a "hydrophobic vacuum cleaner" binding non polar, hydrophobic compounds partitioned into the membrane and subsequently discharging them into the extracellular medium. Nowadays evidence points more in the direction of a drug "flippase", moving the substrates from the cytoplasmic membrane leaflet to the extracellular

leaflet where they can partition into the aqueous phase[35] (Figure 1.4).

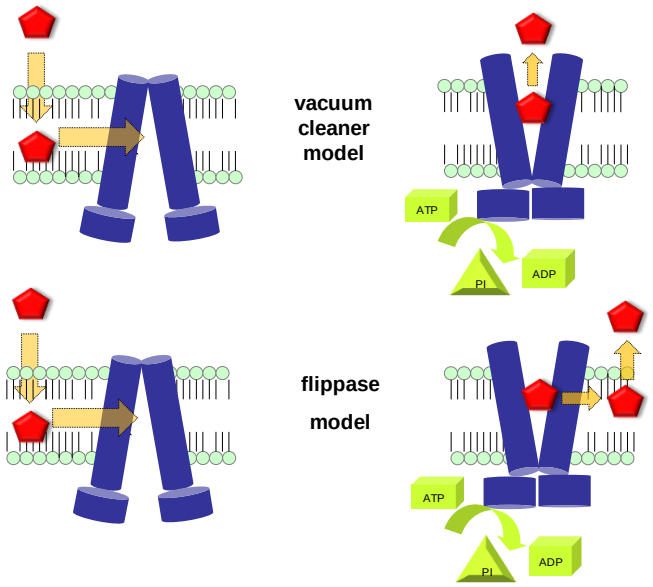


Figure 1.4: Schematic depiction of vacuum cleaner model and flippase model

In Step I of the catalytic cycle the ligand binds to the TMDs in the high affinity open inward conformation and consequently enhances the affinity for ATP. After ligand binding conformational changes in the NBD occur causing the higher binding affinity for ATP.

In Step II ATP binding induces closing of the NBD dimer. An undisputed fact is that ATP binding not hydrolysing presents the energy stroke powerful enough to merit major transformational changes in the TMDs. Another, as important, aspect is the fact that ATP binding and not substrate binding decides which conformation the transporter is about to take. The inward state of the transporter changes to the outward facing state (Sav1866) to the extracellular medium.[39] Of course such a major transformation also changes the binding affinity of the ligand in the cavity which moves from high binding affinity to low binding affinity leaving the ligand to enter the extracellular medium. An ongoing discussion provides the question if two ATP molecules at the same time or one after another can induce the enormous transformational change. The ATP switch model[37] of Higgins and Linton

was the first model correctly claiming the responsibility of ATP binding opposed to drug binding. The authors postulated that ATP binding confers the directionality, kinetic advantage and energy for substrate transport and assume that in the closed dimer both nucleotide-binding pockets have bound ATP. They are challenged by the "occlusion-induced switch model". This hypothesis suggests that two ATP hydrolysis steps are required instead of one. First comes the occlusion and hydrolysis of the first ATP for the high to low affinity transition followed by the second ATP for resetting the transporter to its starting position.[40] Recently it was proposed that ABCB1 exists in a so-called asymmetric state where one active site would be open with low substrate affinity and the other one closed with high substrate affinity and the two NBDs would alternate between ATP hydrolysis.[17]

Step III provides ATP hydrolysis which then destabilises the closed NBD dimer and initiates return into starting position.

Step IV further means the release of Pi and then ADP in completion of the transport cycle. The protein is then once again ready in a high binding-affinity state for ligands. As affinity of a transporter to ADP is low it is unable to stabilise the NBDs presenting a bit of a question mark to the published structure of Sav1866 crystallised with ADP[37] (Figure 1.5).

Ernst *et al.*[41] introduced the kinetic substrate selection model. They propose that both the kinetics of transporter-substrate and transporter-nucleotide interactions affect the substrate selectivity. As explained above ATP binding seems to be the rate-limiting step between inward and outward conformation and transport of ligands. They bring into play two drugs, one called FAST and another one SLOW which display different kinetics. Drug FAST has fast on and off kinetics and drug SLOW shows slow on and off kinetics. At a given time the transporter switches to its starting position and both drugs start to bind into the cavity. Therefore the rate at which the transporter switches back to the outward-facing conformation determines which of the two substrates will be transported. Either it exports the drug FAST very efficiently but no drug SLOW or both drugs but not as efficiently as before. This aspect shows again that as soon as some questions,

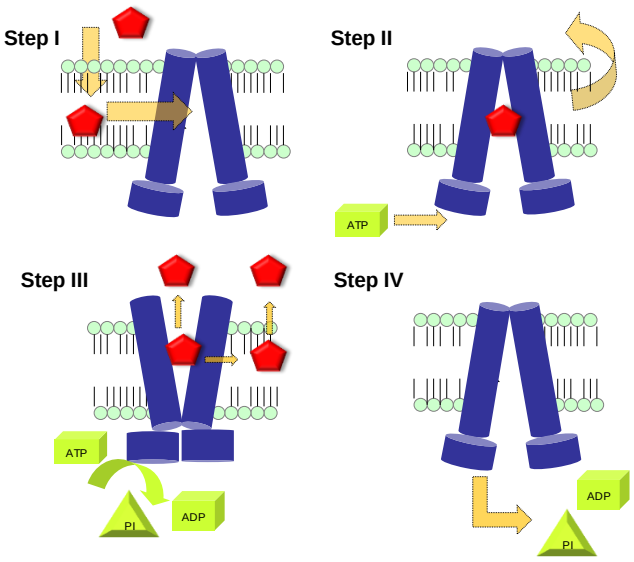


Figure 1.5: Schematic depiction of catalytic cycle: Step I – substrate binding, Step II – ATP binding and conformational change, Step III – ATP hydrolysis and substrate expulsion, Step IV – release of ADP and P_i

like crystal structures, are partially answered new ones immediately turn up rendering understanding of the ABCB1 transporter function and action still exceedingly difficult.

1.2.6 Homology models

Recently John Wise published a series of molecular dynamics simulations of the catalytic cycle based on a homology model of the partially opened outward Sav1866 and docking studies with substrates and inhibitors of ABCB1.[42] In principle the molecular dynamics simulations confirmed the theories postulated by Loo *et al.*[27] of a substrate-induced fit mechanism but interestingly the docking studies did not reveal any special binding site for modulators/inhibitors other than for substrates. The flexibility of the drug binding domain was further supported by another molecular dynamics study by Liu and coworkers[43] and especially the flexible transmembrane helices 6 and 12 seem to be essential for the process of ligand binding whereas transmembrane helix

4 and 5 render the transporter stable. After ATP hydrolysis conformational changes take place. The high affinity residues in the drug binding pocket relocate and substrate affinity decreases rapidly.

The fact that many studies have been rendered unusable after the retraction of the original crystal structure by Chang and colleagues shows the uncertainness of structure based approaches. Sav1866, the MsbA structure and mouse Pgp have provided us again with more or less suitable templates for homology modelling. Nevertheless homology models suffer from interpretive limitations in crystal structures and it remains extremely important to carefully validate the X-ray structure. Also cross link studies and ligand phtoaffinity labelling could be interpreted in different ways and convincing hypotheses may still come to nothing. For example the sequence identity in the TMDs covers only about 20% identity and modelling is done in the so called "twilight zone".[24,44]

1.3 Biological assays for multidrug resistance ABC transporter substrate properties

As the definition of substrate or non-substrate properties of a compound is the basis for every in silico prediction study enormous importance lies herein. In order to highlight the difficulties in achieving these properties an overview over biological assays used to characterise ABCB1 substrates is provided in this chapter. In pharmaceutical industry more and more companies require the means for a cost efficient automatic screening of ABC transporter substrate properties in order to predict possible problems concerning oral bioavailability, adverse central side effects and drug drug interactions.

1.3.1 *In vivo* models

Of course *in vivo* model systems present a valuable tool to evaluate the role of ABC transporters. They mostly are generated with MDR1 knock-out mice comparing their blood/brain exposure ratios to the wild type mice. However, those tests involve enormous expense considering the time effort and the small

number of compounds actually tested. Also it must not be forgotten that factors like solubility, plasma protein binding, passive diffusion and active transport merge into a combined result and do not specifically describe ABC transporter export. Hence input of these results in computational studies is not feasible.

1.3.2 *In vitro* models

Normally *in vitro* assays use cells stably or transiently over expressing MDR-ABC proteins, or membranes/proteins isolated from these cells. Preferable for these assays are direct set-ups where the substrate pathway can be easily followed. Regrettably, this is not always possible as ABC substrates mostly are hydrophobic and diffuse into the cell membrane. Important factors like the mentioned hydrophobicity, availability of binding sites and permeability of the cell membrane demand the use of indirect methods. Here the distinction between substrate or inhibitor provides some difficulties rendering assignation of clear substrate properties sometimes questionable.[4]

Another aspect of the problem is a universally agreed nomenclature of what exactly and with which properties a compound is pronounced substrate, non-substrate, inhibitor and modulator (competitive substrate). A substrate binds to the transporter and then is extruded from the inner cell (low intracellular accumulation and high ATPase activity). An inhibitor blocks the extrusion of a substrate (high intracellular accumulation and no ATPase activity) and a competitive substrate, from now on called modulator, prevents the extrusion of a substrate by saturating the transporter (high intracellular accumulation of original substrate and high ATPase activity). A non-substrate on the other hand passively diffuses through the membrane and never shows any interaction with the transporter (high intracellular accumulation and no ATPase increase). A substrate whose passive diffusion rate is very fast may appear as non-substrate whereas it is exported from the cell by the ABCB1 transporter though rapidly diffuses back into the cell. Here one would have to measure (high intracellular accumulation and high ATPase activity).[45] (Figure 1.6).

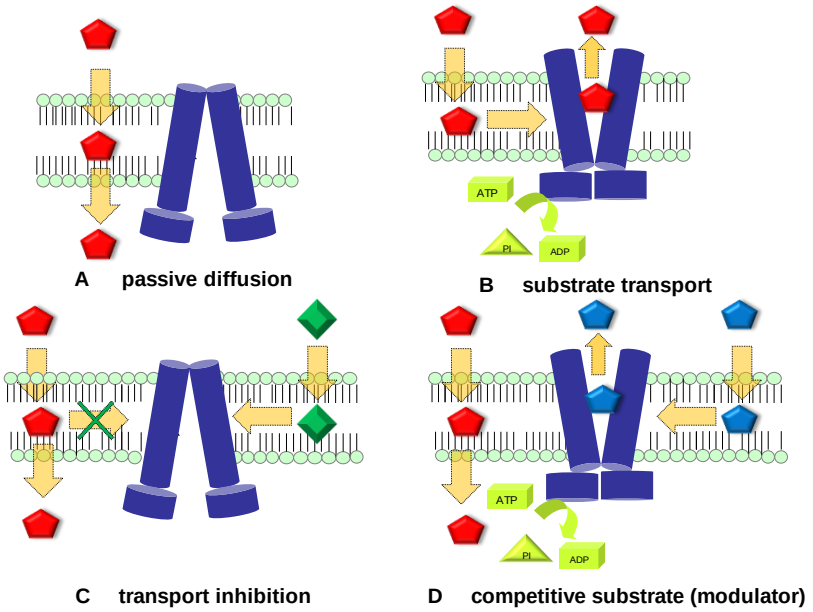


Figure 1.6: Differences between A — non-substrates, B — substrates, C — inhibitors, D -– modulators (competitive substrates).

1.3.2.1 Cytotoxicity

Here stable cell lines overexpressing the corresponding ABC transporter are exposed to increasing levels of cytotoxic compounds. A count of the surviving cells, establishing the respective IC_{50} (concentration where the growth of the cells is inhibited by 50.[46] Zhou *et al.*[47] measured cytotoxic effects on a panel of drug-sensitive (H460) cancer cells and paclitaxel-resistant (H460taxR) cell lines with the ambition to discover novel cytotoxic non-substrates of ABCB1.

1.3.2.2 Intracellular accumulation

Based on ABC function high accumulation of the compound in question inside the cell proves that no export has taken place whereas low intracellular accumulation shows export out of the cell. If the compound in question has intrinsic fluorescence or a radioactive label direct measurement can be used. Otherwise a fluorescent reference substance for ABCB1 namely calcein-AM is mainly employed. Calcein-AM enters the cell by passive diffusion and then is exported by ABCB1 out of the cell. Whenever it competes with another

substrate transport is hindered and high intracellular concentration can be measured. In case of non-substrates extracellular level of calcein-AM is high. However, the indirect method makes it impossible to differentiate between inhibitors and non-substrates and toxic test compounds may result in false measurements. This type of assay can also be transformed in an inhibitor assay. Here one can distinguish between inhibitors and non-inhibitors by measuring ABCB1 efflux of the fluorescent calcein. However, it is important to be aware that in essence this assay is more suited for ABCB1 inhibitors and if used for ABCB1 substrates no correlation between substrates and inhibitors may exist.[4,48]

1.3.2.3 Transport (Cellular monolayer efflux assay)

Cellular monolayer efflux assays more or less mimic ADME (absorption, distribution, metabolism, elimination) effects *in vivo*. Here cellular monolayer efflux enables the quantification of direct transporter activity and permeability of the compound through polarized cell monolayers. The test compound is added into the apical or the basolateral solution. At a given time both respective concentrations are measured. The apical (towards inside of cell) to basolateral (towards outside of cell) (A-B) and basolateral to apical (B-A) permeability is then determined. The resulting ratio of the two transport rates allows direct conclusions respective substrate properties. The International Transporter Consortium[15] recommend a ratio basolateral to apical (B-A) to apical-basolateral (A-B) of at least 2 for possible substrates of ABCB1. In this case testing takes place a second time with an established inhibitor and if the compound export is inhibited the test compound can be pronounced substrate. Otherwise non-substrate status has to be given. For indirect use of the assay known reference substrates and known inhibitors are used. Cell lines frequently taken for this assay include human colonic adenocarcinoma (Caco-2), Madin Darby canine kidney (MDCK) and porcine kidney epithelial (LLC-PK1) cells. These measurements of Caco-2 cells are thought to model absorption through the gut wall. However, certain constraints have to be applied. The compounds must not damage the cells themselves and

should diffuse moderately through the cell layer. High passive transport will falsify results as rapid passive diffusion occludes the functions of the ABCB1 transporter.[4,49] Highly permeable compounds may lead to insufficient drug concentrations within the inner membrane as well as 100% use of ABCB1 transport function. In these cases Fluorescence assays and ATPase activity assays are of better use with preference given to Fluorescence assays as they can also be used for identification of ABCB1 inhibitors. Nevertheless in "real life" conditions compounds with such a high permeability rates are the exception and not the rule therefore the monolayer efflux assay nevertheless is the gold standard assay, also according to the International Transporter Consortium.[15] This assay is considered the best *in vitro* model for *in vivo* ADME interactions, i.e. gut absorption, CNS distribution and hepatic or renal excretion. Furthermore, it is also possible to distinguish modulators from inhibitors by testing transport. Disadvantages are its labor intensity and the fact that polarized efflux cell lines often express many other uptake and efflux transporters which can falsify results. Also, levels of ABCB1 expressions vary from lab to lab thereby hindering straight comparability of results.[18]

1.3.2.4 ATPase activity measurements

As discussed in the previous chapter ATP binding provides the power stroke for transport of MDR ABC proteins and is subsequently hydrolysed further on in the catalytic cycle. Therefore used ATP means enhanced transport action has taken place. In this setup ATPase activity can be quantified by measuring ATP consumption, ADP release or the liberation of the inorganic phosphate. With function of the ABC exporter substrates increase the rate of ATP binding and consequent hydrolysis. Whereas ABC inhibitors lower the rate of normal ATP consumption. In case of ABCB1 some substrates showed inverse concentration dependent ATPase stimulation, i.e. low concentrations increase ATPase activity and high concentrations decrease ATP hydrolysis. However, even if no transport action takes place a basal rate of ATP uptake rate is seen due to the lipid environment in experimental

conditions, partial uncoupling or endogenous stimulation. This assay can be used for high-throughput-screening, investigating drug-drug interactions and kinetic issues. Nevertheless, a major disadvantage lies in the interpretation of results which is not always easy. Due to basal ATPase activity low affinity substrates may not be detected with this method. In general this assay is not particularly suitable for differentiation between substrates and inhibitors and is not routinely used as drug screening assay. Nonetheless it provides the opportunity of testing specific drug-transporter interactions.

1.3.2.5 Photoaffinity labelling

In this case the binding of compounds to the transporter is the piece of interest. Photoaffinity-labeled compounds are exposed to membranes full of ABC transporters and via UV irradiation the photoaffinity label is permanently attached to the binding site on the transporter. This setup is suitable for high-throughput drug-drug interaction screening whereas the fact that inhibitors and substrates cannot be told apart is a severe drawback.[50]

1.3.2.6 Vesicular transport

For this setup inside-out membrane vesicles prepared from cells are used and the accumulation of the test compound quantified. Inside-out membranes trick the transporter as the NBDs which normally are localised in the intracellular compartment, in this case open to area accessible for compound testing. Analysed substrates therefore are transported into the vesicles rather than out of them. The membranes are then separated from the incubation solution and high sensitivity analytical methods employed for measurement of transported unlabelled molecules. In case of radio or fluorescent label radioactivity is quantified. A drawback of the labelling method is the fact that labeled compounds are not always readily available. Vesicular leakage also presents a problem as hydrophobic molecules may disappear into the membrane and not necessarily be bound to a transporter. Often the inside-out membrane vesicles are used in combination with ATPase activity measuring assays. Nervi *et al.*. pointed out that in cells substrates and inhibitors are

required to diffuse into the membrane before interaction with ABCB1 takes place, building a two-step model, whereas in inside-out vesicles the diffusion step lacks completely. As a consequence actual concentrations for inhibitors to be active may differ considerably from those measured with inside-out vesicles. The group compared ABCB1 ATPase activity in inside-out plasma membrane vesicles to living cells and wanted to explore the importance of intrinsic permeability opposed to ABCB1 efflux rate.[51]

The previous chapter has highlighted the complexity of ABCB1 functions and therefore measurement of compound properties meets with many challenges. Often categories for substrate properties differ from laboratory to laboratory as experimental conditions and interpretations of results are not necessarily similar. These facts place a higher level of insecurity in the measured results which are then input in computational studies.[45] Of course success for every study relies considerably on the quality of the available data. Another reason for different assignations of compounds are simple faults in publications that happen customarily though great care is taken by the responsible authors.

1.4 In silico studies of ABCB1

As described in the previous chapter crystal structures of transporter membranes are hard to achieve and until recently[26] no reliable structure of ABCB1 was available. Even with the crystallized transmembrane proteins still many question marks exist concerning substrate recognition, ligand binding, binding domains, ligand translocation, etc. which still renders structure based methods for ABCB1 difficult and maybe unreliable. Though the transporter proteins of the same family show the same transmembrane domaine property and therefore homology modelling is possible the sequences of the TMDs responsible for ligand interaction are not as highly conserved as the NBDs.[52]

Consequently ligand based methods have been employed more and more. They work upon the principle that different ligands being active on the same target protein must have some structural similarities in order to be able to

interact with the target protein binding site. Full prior knowledge of the target transporter protein is not required as they correlate biological activity (substrates, inhibitors, non-substrates) with physico-chemical, quantum-chemical, electrostatical and other properties calculated and measured as molecular descriptor values. Using these properties combined with the activity profile models are built by applying quantitative structure activity relationship methods and machine learning algorithms. They have to be carefully validated in order to be able to predict the activity profile of other chemical compounds as substrates, inhibitors or non-substrates of the respective target protein. They also may provide insights into structural properties necessary for ABCB1 transport interaction.

Inspired by the impact of Lipinski's simple rule of Five[53] in drug development Gleeson in 2008[54] presented a set of simple rules of thumbs consisting of just four parameters namely molecular weight, log P, positive ionizable and negative ionizable. Driven by the idea that simple parameters though maybe not 100% correct for every specific case are easier to follow and to implement than more complex parameters not as easy to understand he formulated simple rules for ADMET parameters, namely absorption, solubility, permeability, bioavailability, distribution, CNS penetration, plasma protein binding, brain tissue binding, metabolism and toxicity. Based on a dataset of 1975 compounds for ABCB1 he correlated molecular weight >400 with ABCB1 transport though that poses the additional problem that higher weight is accompanied by lower permeability and therefore lower bioavailability. Further general conclusions he drew consist of: Higher lipophilicity >4 was correlated with higher ABCB1 efflux probability. Basic and neutral molecules were more susceptible to ABCB1 export followed by zwitterionic molecules and lastly acidic molecules.

Recently Hitchcock proposed a series of structural modifications for substrates to change ABCB1 efflux rates.[55] Substrate properties of ABCB1 include hydrogen bond donors, primary amides and N-heterocycles with uncapped NH groups among others and according to him either by removing OH

or NH groups or cloaking hydrogen bond donors via a neighbouring hydrogen bond acceptor group substrate status can be changed. Minimisation of ABCB1 contact can be achieved by keeping the hydrogen bond donor count below 2 and the topological polar surface area below 90 Å. Though these statements are highly general and some compounds may act as substrates nevertheless they represent at least guidelines for further consideration.

With an eye on ABCB1 modulators Pearce and colleagues[56] took a series of reserpine analogs operating as modulators of ABCB1 and proposed common properties as essential for binding to ABCB1. They postulated that the common pharmacophore of ABCB1 modulators requires two planar aromatic domains and a basic nitrogen atom. Our group[57] in 1999 could show that the importance of the nitrogen atom depends on its being a hydrogen-bond acceptor rather than a positive ionizable entity.

Based on a set of 100 compounds Anna Seelig[58] proposed a general pattern for ABCB1 substrate recognition. The set was split into substrates (64 compounds), borderline substrates (11), inducers of MDR1 (18 compounds) and non-substrates (7 compounds). As borderline substrates compounds with a very low substrate activity were defined. She differentiated between two types of electron donor patterns, type I containing two electron donor groups with spatial separation of 2.5 ± 0.3 Å and type II consisting of either three electron donor groups separated by 2.5 ± 0.3 Å from each other, with a spatial separation of the outer two donor groups of 4.6 ± 0.6 Å, or by only two electron donor groups with a spatial separation 4.6 ± 0.6 Å. ABCB1 substrates carry on average two type I units whereas ABCB1 non-substrates contain generally no type I or type II units. In a later study she and coworkers[59] explored the permeability of a substrate at its rate-limiting step for interaction with ABCB1. In direct competition between two substrates the substrate with higher capacity and strength of forming hydrogen-bonds is preferred for transport.[45]

1.4.1 Pharmacophore models for ABCB1

A pharmacophore is the representation of the spatial arrangement of structural features that are required for a certain biological activity.[52] By providing a template for structural features they guide the medicinal chemist in the right direction depending on whether he wants to create substrates, inhibitors or non-substrates. Many program packages such as Catalyst, LigandScout, Phase, MOE offer the means for these projects.[60] In a recent review Chang and colleagues[61] provided a comprehensive account of pharmacophore based discovery of ligands for drug transporters highlighting also virtual screening. They describe two methods available for computational design of new ligand compounds, namely the *de novo* approach and the virtual screening approach. In the *de novo* approach a quantitative model is generated and subsequently guides the further synthesis options of the medicinal chemist. In the virtual screening approach the readily available compounds from commercial vendors in large databases are searched for new leads according to special criteria, for example pharmacophore models. This approach naturally is more cost-effective and cheaper and therefore more tolerable of false positive results in model predictions compared to chemical synthesis.

In 2002 Penzotti and colleagues[62] took the ABCB1 substrates and MDR1 inducing compounds from Seelig's dataset. If transport had taken place they were regarded as substrates and if not non-substrate status was given. Other compounds from literature sources were added to these compounds resulting in a dataset of 195 compounds. This set was further split into training and test set and consequently an ensemble model of 100 pharmacophores established. The pharmacophores consisted of two-, three- and four-point pharmacophores able to differentiate between ABCB1 substrates and non-substrates. By ranking the pharmacophores obtained from each substrate and non-substrate in the training set the 100 best discriminating pharmacophores were transferred into the ensemble model. Compounds that matched at least 20 of the 100 pharmacophores in the ensemble model were considered to be substrates. Results confirmed the type I or type II recog-

nition patterns postulated by Seelig. Though the training set showed an overall prediction accuracy of 80% the overall external accuracy for the test set achieved only 63% with 53% accuracy on substrates and 21% accuracy on non-substrates.[45]

Though focusing primarily on ABCB1 inhibitors[63] Ekins *et al.*[64] also established a pharmacophore based on verapamil, vinblastine and digoxin. They confirmed that the modulator verapamil is likely to bind on the same site as vinblastine/digoxin on ABCB1. Their pharmacophore showed identical features like hydrophobic regions, hydrogen bond acceptors and aromatic ring properties and seemed to partly overlap with the four inhibitor pharmacophores they formerly created. Also, they detected some similarity in comparison with other studies on CYP3A4 inhibitors.[65,66] The pharmacophores of both proteins contain multiple hydrophobic features and at least one hydrogen bond acceptor though in slightly different arrangements. This finding once again shows the high similarity between CYP3A4 and ABCB1. In a follow-up of their recent publication[67] Chang and colleagues performed virtual screening of the Penzotti dataset[62] on the three pharmacophores formerly obtained[63,64] followed by *in vitro* testing of the in-house SCUT database to assess the prediction accuracy. One pharmacophore model for substrates and two pharmacophore models for inhibitors were developed (Figure 1.8). Also, two different searching methods FAST (rigid) and BEST (flexible) were employed. In concordance with other studies the substrate model retrieved compounds with a higher number of hydrogen-bond acceptors and donors and also compounds with higher molecular weight. Search method FAST resulted in the highest enrichment factor of 1.8 for the substrate model. Four of six returned drugs of the in-house database were confirmed as known ABCB1 substrates or inhibitors.[45]

Pajeva and Wiese[68] proposed a general pharmacophore for ABCB1 substrates and modulators. It consists of two hydrophobic, three hydrogen bond acceptor and one hydrogen bond donor and was specifically restricted to diverse substrates of the verapamil binding site. Based on their results

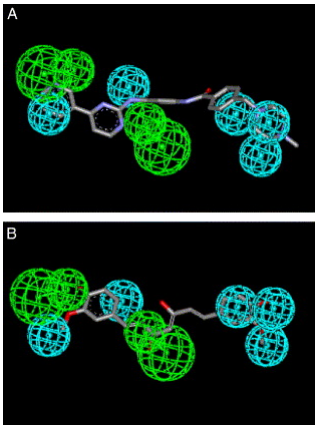


Figure 1.7: Pharmacophore model; taken from Chang *et al.*[61] Copyright Advanced Drug Delivery Reviews 2006. Catalyst P-gp substrate pharmacophore with (A) Gleevec, (B) curcumin mapped to hydrophobic (blue), and H-bond acceptor (green) features.

they postulated two hypotheses in order to explain the structural variety of ABCB1 substrates: First that the verapamil binding site of ABCB1 has several participants in hydrophobic and hydrogen bonding interactions and second, that different drugs can interact with different receptor points in different binding modes. In comparison to the study on ABCB1 inhibitors by Ekins and colleagues[63] also targeting the verapamil binding site some differences concerning the distance matrices between the presented pharmacophoric features are evident. This once more highlights the flexibility and promiscuity and challenge that goes with targeting the ABCB1 transporter protein.[45]

In 2002 Garrigues and colleagues[69] also presented two pharmacophores based on ABCB1. During *in vitro* testing they confirmed that some compounds in their data set showed non-competitive interactions with other molecules thereby proving definitely that at least two different binding sites on ABCB1 exist. Other molecules on the other hand showed competitive binding interaction. The authors suggested either binding on the same site or steric constraints that hinder simultaneous binding by large, rigid molecules. Their two pharmacophores were based on three dimensional consensus

hydrophobic and polar elements and partially overlapped. They also showed a correlation between molecular weight and binding affinity to ABCB1 maybe due to the larger number of interaction points provided. Substrate recognition according to Garrigues takes place at defined sites, taking into account the shape, size and distribution of hydrophobic and polar elements of substrates. Thereby the recognition of various chemical structures is made possible.[45]

In another study Cianchetta *et al.*[70] used an approach combining physicochemical properties of a molecule (Volsurf Descriptors) and pharmacophoric properties (GRIND descriptors obtained with Almond). GRID-alignment-independent descriptors (GRIND) are able to describe pharmacodynamic properties based on 3D conformations but independent of their position in space. They also contain shape descriptors, describing the distance between certain regions by spatial extent of the molecule. A dataset of 129 compounds was used and split into training (109 compounds) and test set (20 compounds) and partial least square multivariate data analysis (PLS) combined with the feature selection algorithm FFD employed. In the resulting model the contribution of the pharmacophoric descriptors seemed to be more significant than that of the physicochemical descriptors. The most important physicochemical descriptors were size and shape and the hydrogen-bonding capabilities. On the whole the pharmacophoric descriptors were in general accordance with previously found pharmacophoric features, consisting of two hydrogen-bond acceptor groups and two hydrophobic areas. They also emphasised the influence of the size of the compounds to substrate activity.[45]

Crivori and colleagues in 2006[71] first developed a computational model based on Volsurf descriptors and partial-least-squares discriminant analysis (PLSD). Second 3D pharmacophores based on the aforementioned GRIND descriptors using Almond were calculated in order to differentiate between substrates and not transported inhibitors. The dataset consisted of 53 substrates and non-substrates assigned after *in vitro* testing with the transport assay. The model based on Volsurf descriptors showed an overall accuracy

of 89% for the training set and for the 257 compounds comprising test set 72% overall accuracy and 61 % accuracy on substrates and 81% accuracy on non-substrates was achieved. The most important features of the compounds consisted of size, shape and flexibility of the molecules. For the approach discriminating between substrates and inhibitors the 53 drugs were tested *in vitro* and nine ABCB1 substrates and 14 ABCB1 inhibitors taken for modeling. GRIND descriptors were calculated and PLSD was used and the resulting model evaluated with a test set consisting of 68 substrates and 56 inhibitors taken from literature. The model obtained an overall accuracy of 82% with 88% accuracy on substrates and 75% accuracy on inhibitors. Inhibitors should contain favourable interaction regions placed 8 Å apart around two hydrogen bond acceptor groups and a hydrophobic region and a hydrogen bond acceptor group separated by a distance of 4.0 Å.

Li *et al.*[72] showed an impressive performance with the combination of a classification study based on many pharmacophores as descriptors. They achieved an overall accuracy of 87.7% for the training set of 163 compounds, taken from Penzotti[62] and 87.6% for the test set of 97 compounds, taken from literature. Their method started with an exhaustive search of all possible pharmacophore hypotheses for both substrate and nonsubstrate compounds resulting in 12.6 million possible pharmacophores. Further on they identified a statistically significant optimal pharmacophore ensemble with the ability of substrate/non-substrate discrimination. This was achieved by employing a frequency analysis counting the most frequent pharmacophore with a minimal pharmacophore occurrence cutoff though no maximum cutoff was necessary. The remaining approximately 63000 potential pharmacophores were evaluated with a pharmacophore specific t-statistic score resulting in an optimal set of 598 pharmacophores. The important pharmacophoric features identified in this study back up the findings of other studies as the most important features consist of aromatic rings, hydrophobic areas and hydrogen bond acceptors. Interestingly hydrogen bond donor features occurred only in 28% thus supporting the suggestion that ligand acceptor interactions are the most significant. Also the type II-patterns described by Seelig[58] were

a common occurrence opposed to the type I pattern which appeared seldom. In the next step the number of 598 pharmacophores was diminished by recursive partitioning and the best resulting decision tree contained only the nine top ranked significant pharmacophores. Furthermore, each of the nine significant pharmacophores can be used as standalone. Five of them showed accuracy on substrates of 1.00 with the worst performing model achieving 0.78. This combination of many pharmacophore models with a classification algorithm successfully brings together the best of two worlds and represents a successful state-of-the-art method.

Recently another study regarding ABCB1 substrates has been published.[73] The authors performed flexible receptor docking based on the crystallized mouse P-glycoprotein structure and further docked 102 metabolites into the binding pocket with an area under the curve of 0.93 based on a ROC type curve. Taking a dataset based on FDA approved drugs and respective assays they tried to discriminate between 13 substrates and 34 non-substrates with a resulting area under the curve of 0.90. They state their belief that more success of understanding the specificity of ABCB1 substrate recognition can be found by regarding ligand properties than by the big drug binding pocket of this still mysterious protein.

Though some features like hydrophobic, hydrogen bond acceptor, hydrogen bond donor properties repeatedly are mentioned in the studies presented direct comparison of those models is not possible due to the differences in the distance matrices of those features. This circumstance once again highlights the promiscuity of ABCB1 and suggests that for the highly diverse substrate profile of ABCB1 ensemble methods as employed by Penzotti and Li achieve best results. However in recent years several studies have been conducted to define chemical features commonly shared by ABCB1 substrates (Table 1.1). Contrary to pharmacophore models, these methods define rules that discriminate between substrates and non-substrates based on calculated descriptor values reflecting physico-chemical properties.[45]

study	N	performance		features
		training	validation	
Penzotti *et al.*, 2002[62]	195	Acc: 80%	Acc: 63%	ensemble PH4: 53 four-point; 39 three-point; 8 two-point
Ekins *et al.*, 2002a[63]	16	r^2=0.96	r^2=0.72 (RG)	2 HYD, 1HBA, 1Ar
Pajeva and Wiese, 2002[68]	20	n.a.	n.a.	2 HYD, 3HBA, 1HBD
Ekins *et al.*, 2002b[64]	27	r^2=0.77	r^2=0.43	4 HYD, 1HBA
	21	r^2=0.88	r^2=0.31	3 Ar, 1HYD
	17	r^2=0.86	r^2=0.64	2 HYD, 1HBA, 1HBD
Garrigues *et al.*, 2002[69]	n.a.	n.a.	n.a.	PH1: 1 Ar, 2 Alkyl, 1e–don PH2: 1 Ar, 3 Alkyl, 1e–don
Cianchetta *et al.*, 2005[70]	129	r^2=0.81	q^2=0.72 (RG)	2 HYD, 2HBA, size
Chang *et al.*, 2006[67]	33	r=0.87	r=0.56 (RG)	4 HYD, 1HBA
Crivori *et al.*, 2006[71]	53	Acc: 88.7%	Acc: 72.4 %	HBA, HBD, HYD, distance
Crivori *et al.*, 2006 (inhibitors)	23	n.a.	Acc: 82.4%	1 HBD, 3 HBA, hydrophobic
Li *et al.*, 2007[72]	163	Acc:87.7%	Acc: 87.6%	9 models out of 12.6 million four-point PH4

Table 1.1: Overview over pharmacophore studies performed. Taken and adapted from Demel *et al.*[45] Copyright Future Medicinal Chemistry. PH4 – pharmacophore, HYD – hydrophobic areas, HBA – hydrogen bond acceptors, HBD – hydrogen bond donors, Ar – Aromatic features, Alkyl – alcylic chains, e—don – electron donor, Acc – overall prediction. accuracy

1.4.2 Classification studies for substrate prediction

In 2003 Didziapetris *et al.*[74]published a classification analysis of ABCB1 substrates and non-substrates thereby presenting the rules of four. They stressed that 3D modelling approaches are sometimes questionable for highly dissimilar compounds as alignment can be very challenging. In their study they used two datasets derived from earlier studies. The first dataset contained 220 compounds, mostly drugs, drug candidates or natural products obtained from literature. The second dataset consisted of the 220 compounds of dataset I and 780 additional compounds also retrieved from literature and was only used as a clarification tool. After calculation of simple ADME descriptors like number of H-accepting and H-donating atoms, strongest acidic group, strongest basic group, number of aromatic rings, etc these descriptors were analysed using recursive partitioning analyses. Similar to Gleeson[54] later the results of that analysis gave them leave to postulate simple rules for ABCB1 substrate and non-substrate properties. The "rule of fours" as proposed by the group, says that compounds with (N + O) $\geq$ 8, MW $>$ 400 and acid $p\mathrm{K}_a > 4$ are likely to be substrates whereas compounds with (N + O) $\leq$ 4, MW $<$ 400 and basic $p\mathrm{K}_a < 8$ are likely to be non-substrates. Secondly Didziapetris *et al.* expressed their conviction that ABCB1 acts as a "small chromatographic pump" and used Abraham's solvation equation for ABCB1 substrates. The large binding site of ABCB1 was thought to neglect any 3D conformational effects. Thirdly they presented class-specific models, providing specific substructures commonly found in certain ABCB1 substrates. Based on these data the authors implemented an ABCB1 substrate specificity prediction tool into their software ADME Boxes thereby enabling the user to classify compounds either as ABCB1 substrates, inhibitors or inconclusive.[75,76] By providing a reliability index the user can clearly see whether his compounds of interest lie in the applicability domain of the ADME Boxes model. In order to roughly categorise one's compounds this tool may give an idea for probable class membership.[45]

In the course of the development of a Blood-Brain Barrier (BBB) perme-

ation model Adenot and Lahana[77] also built a model for ABCB1 substrates and non-substrates as efflux plays an important role in BBB permeation. For this model the substrates of the ABCB1 dataset from Anna Seelig[58] were taken, consisting of 91 compounds. As non-substrates 1545 compounds were used, obtained from the WDI (world drug index) but no inducers or modulators were allowed except inhibitors which are substrates at the same site as Verapamil. However, some of the non-substrates in the training set have a partly undefined status respective their real interaction pattern with ABCB1 thereby some level of insecurity remains. For the ABCB1 test set a sub-selection of the Seelig compounds was chosen with 13 ABCB1 non-substrates and 20 ABCB1 substrates. Relatively simple 2D and 3D derived properties like ADME, topology, electronic, energy, surface and geometric descriptors were calculated and further linear discriminant analysis (LDA) or partial least square discriminant analysis (PLS-DA) employed. The LDA model did not show very promising results whereas PLS-DA was more satisfactory. A sensitivity of 70% for ABCB1 substrates and a specificity of 92% for ABCB1 non-substrates could be achieved. As the ABCB1 substrate compounds lowered classification rate for non-crossers in the BBB model they were excluded from the final model. As a consequence seemingly impermeable compounds, such as cyclosporin A become highly permeable when ABCB1 is saturated. This highlights once again the high importance of ABCB1 as protector of the central nervous system.[45]

In 2004 Gombar and colleagues[78] published a study of 95 compounds, all assayed by one monolayer efflux assay based on the MDR1-MDCK cell line. They calculated electro-topological descriptors, shape descriptors, molecular weight, molar refraction, hydrogen bonding donors and acceptors and lipophilicity. After obtaining a model with a two group linear discriminant function they received an overall external accuracy of 86,2%, a sensitivity of 94,3% and a specificity of 78,3% for a test set of 58 additional compounds. The model was based on the remaining 27, of formerly 254, descriptors reduced by stepwise discriminant analysis and enabled the authors to state the Gombar-Polli-E-state Rule. This rule is based on the relationship between

the molecular bulk and ABCB1 class membership. Compounds with MolES > 110 were substrates and the majority of compounds with MolES < 49 consisted of non-substrates. Also they once again stress the enormous time and cost lost over assays for assignation of ABCB1 substrate or non-substrate status. For approximately 100 compounds two full-time employees had to spend 2 months doing the experimental and analytical work necessary. Seen on a large scale quick and easy classification methods represent an enormous advantage in screening possible ABCB1 substrates or non-substrates as required.[45]

More sophisticated machine learning methods were used by Xue *et al.*[79] on a data set of 201 compounds further split into training set (142 compounds), testing set (34 compounds) and independent validation set (25 compounds). The 159 descriptors encompass simple molecular properties, molecular connectivity and shape, electro-topological state, quantum chemical properties and geometrical properties. These were further reduced by a feature selection method, in this case recursive feature elimination (RFE) followed by a support vector machine approach using a Gaussian kernel function. A support vector machine tries to find the best hyperplane able to separate objects in a multidimensional feature space and can be optimised by parameter selection. In order to compare model performances the authors also employed k-nearest neighbour (kNN), probabilistic neural network (PNN) and C4.5 decision tree. These four methods were directly compared during five-fold cross-validation. The kNN approach achieved 70,8% overall accuracy, the neural network resulted in 74,4%, the decision tree in 71,5% overall accuracy and the support vector machine approach presented the best overall prediction in five-fold cross-validation with 79,4% accuracy. For the independent test set this method achieved an accuracy on substrates of 84,2% and an accuracy of non-substrates of 66,7%. However, as the independent validation set comprises only six non-substrates opposed to 19 substrates these results are slightly biased. Following the success of the support vector machine approach compared to other methods the authors investigated the direct merit of the feature selection method (RFE). On three data sets, the ABCB1 data

set already used, a data set of human intestinal absorption molecules (HIA) and compounds causing torsade de pointes (TdP)[80] they used the same approach as before with and without feature selection. Xue *et al.* could show a significant increase in model accuracy due to feature selection. For ABCB1 the accuracy on substrates increased from 68,9% to 81,2% and the accuracy on non-substrates from 68,2% to 79,2% by using five-fold-cross-validation. The overall cross-validated accuracy improved from 68,3% to 79,4%. For presentation of their results they used the Matthews Correlation Coefficient (MCC) which is a discrete version of Pearson's correlation coefficient usable even for classes of very different sizes.[45]

Taking up the idea of neural networks Wang *et al.*[81] used both the supervised back propagation neural network (BPNN) and unsupervised Kohonen self-organising maps (SOM). Taking a data set of 206 compounds (substrates and inhibitors) compiled from literature 248 electro-topological state descriptors (E-State) and molecular connectivity indices were calculated. After the application of feature selection through stepwise discriminant analysis 11 descriptors remained which were fed into SOM and the BPNN. Direct comparison favoured the unsupervised approach considering the substrate diversity of ABCB1. SOM achieved an accuracy on substrates of 83,3% and an accuracy on inhibitors of 80,8% being able to discriminate substrates against inhibitors very well.[45]

Introducing a novel classification method called random feature subset boosting for linear discriminant analysis (LDA) Arodz and colleagues[82] performed their method among others on the set of ABCB1 compounds chosen by Xue *et al.*[79] Their new method involves combining multiple models to obtain one more reliable model. New ensemble members are built in order to correct errors of previously trained members and thereby achieving higher accuracy. As formerly LDA was not susceptible to boosting the new method introduces the concept of randomly chosen subsets of descriptors to boosting similar to the concept of random forest but with another method underneath. With their approach based on only four descriptors they achieved an overall

accuracy for the ABCB1 data set of 80,8% using the same descriptors compared to the overall cross-validated accuracy of Xue *et al.*[79] of 79,4% despite the feature selection method RFE.

Cabrera and colleagues[83] in 2005 introduced their TOPS-MODE (topological substructural molecular design) approach. It is based on the calculation of the spectral moments of the bond matrix. These descriptors measure the degree of concentration of physicochemical properties (hydrophobic/polarity, electronic and steric, charge, van der Waals atomic radii, molar refraction, atomic mass). After compilation of a data set of 203 compounds from literature the data set was split into 163 compounds as training set and 40 compounds for the external test set. In order to prove that no overfitting of the model had taken place the authors subjected the descriptors to Randic's orthogonalisation procedure. The most important descriptors used by the model were standard bond distance describing the molecular size, polarizability and atomic charge. As classification method linear discriminant analysis was employed which achieved an overall external prediction accuracy of 77,50% with a sensitivity of 81,82% and a specificity of 72,22%. The high significance of the bond distance in this context suggests that it could be considered as a more general property for discrimination between ABCB1 substrates and non-substrates. An altogether independent set belonging to the 6-fluoroquinolone family confirmed the reliability of the model with an overall accuracy of 77,7%.

Taking into account the numerous machine learning methods de Cerqueira Lima and colleagues[84] developed a combinatorial QSAR approach employing various optimisation methods and descriptor types. Four methods, k-nearest neighbour (kNN) classification, decision tree (DT), binary QSAR (BQSAR) and support vector machine (SVM), have been employed using the 195 compounds taken from Penzotti *et al.*[62] with the same training and test set. Also four different sets of descriptors were calculated, consisting of molecular connectivity indices, atom pair descriptors, VolSurf descriptors and molecular operating environment (MOE) descriptors. Classification studies were

carried out separately for each method and descriptor type resulting in a total of 16 combinations. The authors introduced a new definition of the correct classification rate (CCR) which is defined as the ratio of correctly classified compounds to the total number of compounds multiplied by 0.5. The best method and descriptor set was obtained by SVM classification in combination with atom pair descriptors or VolSurf descriptors resulting in an external sensitivity of 78% and an external specificity of 84%. All models had higher accuracy in classifying non-substrates than substrates and the authors stress the advantages of combinatorial approaches as one method and one descriptor type are more susceptible to failure.

A very good method for ABCB1 classification was proposed by Huang and colleagues[85] in 2007 resulting in an external overall accuracy of 88,6%. The data set used was taken from Cabrera and colleagues[83] and the same training (163 compounds) and validation set (40 compounds) used. 929 molecular descriptors were calculated and after elimination of redundant information the remaining 79 descriptors subjected to feature selection using a particle swarm algorithm. This algorithm is a stochastic and population based search algorithm where each particle is randomly initialised with an original position and velocity. After descriptor reduction a support vector machine with a Gaussian radial basis function (RBF) kernel after parameter optimisation was employed. The combined efforts yielded an external sensitivity of 82% and an external specificity of 91%. These results are impressing considering that only seven non-correlated and simple descriptors have been used. They encompass three constitutional descriptors, two functional group counts and two molecular property descriptors. As previously suggested molecular mass, hydrogen bonds and polar surface area play major parts in substrate recognition of ABCB1 but in this study features such as the number of ring tertiary C atoms and the number of substituted benzene C atoms seem to be relevant.[45]

In 2007 Zhang *et al.*[86] presented a classification study for efflux substrates based on *in vitro* bidirectional Caco-2 cell permeability. As Caco-2 cells express several efflux transporters such as ABCB1, breast cancer resis-

tance protein, multi-drug resistance protein, etc this classification method is not exclusively utilisable for ABCB1 substrates. The training set consisted of 125 compounds and 46 compounds as test set. The authors employed the recursive partitioning method (RP) in order to develop a classification tree based on five descriptors characterising physicochemical, shape and atom type properties. The simple descriptor types enable the definition of substrate properties such as electron deficient aromatic rings, highly branched compounds and the frequent presence of tertiary nitrogens. The external validation set achieved a sensitivity of 89% on substrates and a specificity of 72%. False *in vitro* identification as non-substrate could be due to high permeation, very little permeation or the transporter could be saturated by high affinity substrates.

Yang and colleagues[87] presented an approach using both feature selection and parameter optimisation simultaneously for the support vector machine. Two parameters are important when using the support vector machine, i.e. the penalty parameter C and the kernel parameter γ which can be optimised using the conjugate gradient method. Feature selection methods are another opportunity to optimise classification methods and in this case a genetic algorithm was employed. The data sets taken from Xue *et al.*[80] show a cross-validated overall accuracy of 85,1%, a sensitivity of 92,2% and a specificity of 75,3% with 8 of formerly 223 descriptors remaining.

Wang and colleagues[88] collected a data set of 332 compounds consisting of compounds taken from literature, the data set taken from Penzotti[62] and the remaining compounds taken from Seelig.[58] The descriptors used include 2D Autocorrelation descriptors and physico-chemical descriptors obtained from the molecular operating environment (MOE). Three different feature selection methods were employed, namely Pearson correlation analysis and random forest-based feature selection and last the F-score as a measure of feature relevance. For classification the machine learning method support vector machine with RBF-kernel was used. A combinatorial approach was built using each descriptor set with different feature selection methods. The best

model yielded an overall external accuracy of 88% based on 23 descriptors selected via correlation analysis. They describe molar refractivity, lipophilicity, molecular vertex adjacency information and partial charge based on van der Waals surface area.

Taking his experiences of modeling a set of ABCB1 inhibitors with the use of molecular interaction fields in 2011[91] Fabio Broccatelli expanded his studies to ABCB1 substrates.[89] Taking a set of 187 compounds with 110 non-substrates and 77 substrates characterised by one assay a training set, consisting of 150 compounds, and a validation set were selected via principal component analysis. Volsurf descriptors based on molecular interaction fields and 2D descriptors were employed starting from the minimised conformation of each compound and further analysis followed using Orange Canvas and Chembench. After a feature selection procedure models were built based on naïve Bayes, support vector machine, and k-nearest neighbour (kNN) using Orange Canvas and with Chembench random forest, support vector machine and a genetic algorithm for kNN were utilised for model building. Descriptors were selected according to the method used and scored according to their Matthews Correlation coefficient or to their area under the curve (AUC) based on a ROC curve. In this study naïve Bayes always outperformed the kNN models whereas models with descriptors selected based on the Matthews Correlation coefficient always showed better accuracy than when selected via AUC. The best model in this study achieved an external accuracy of 86% and is based on 4 Volsurf descriptors and naïve Bayes as classification method. The other methods of the orange Canvas approach yielded performances of approximately 78% with kNN based on 7 variables and SVM on 11 variables though in the case of SVM no method tailored feature selection took place. With Chembench random forest and the genetic algorithm kNN gave best performances, receiving external accuracies of 84% and 81% respectively. The best models presented in this work were built on Volsurf descriptors and the author emphasised the importance of 3D information for model building.

In 2012 Vasanthanathan Poongavanam of our group produced a predictive model of ABCB1 substrates and inhibitors based on fingerprints.[90] 257 compounds were taken from literature and the same dataset as taken in this study used for model building of ABCB1 substrates. Models were built using WEKA data mining software and were based on random forest, support vector machine and k-nearest neighbour (kNN) achieving an overall external accuracy of 67-70%. Using a rule algorithm, FP Growth, rules for ABCB1 substrates could be derived. These rules suggest hydrophobicity as prerequisite for ABCB1 transport and aromatic systems coupled with ether and amine moieties. Non-substrates on the other hand often contain hydroxyl groups.

Recently Zhang and colleagues have published a study of anti-epileptic drugs and ABCB1 transport.[92] In this field ABCB1 transport also leads to major drug resistance problems and again discrimination between substrates and non-substrates of ABCB1 would enable new therapeutic ways. Based on the published models for ABCB1 transport, especially the report of Seelig, they extracted the most important features and manually assigned transport status for ABCB1 on a set of anti-epileptic drugs. This emphasises again the high need and the higher challenge for discriminative models between substrates and non-substrates of ABCB1.

As proven in this chapter studies of ABCB1 substrate properties present many challenges and throwbacks (Table 1.2). Nevertheless machine learning methods seem to be the most promising approach for reliable ABCB1 substrate prediction. Among the available machine learning methods the support vector machine especially in combination with feature selection algorithms are in general more successful and also provide simple interpretable descriptors.

study	N	algorithm	descriptor types	Acc. (CV)	N (test)	Ext. Acc.	Ext. SE
Seelig[58]	100		hydrogen bonding moieties				
Didziapetris *et al.*[74]	220+1000	C-SAR	ADME Descr. (H acc, pK_a, nrings, TPSA, MW, etc.)				
Xue *et al.*[79]	201	SVM	molecular properties, shape,	79,40	13	n.a.	84,20
	201	kNN	electrotopological state,	70,80			
	201	PNN	quantum chemical properties	74,40			
	201	C4.5 DT	geometrical properties	71,50			
Adenot and Lahana[77]	91	PLS-DA	ADME, geom., top., electr., surface, energy descr.		33		70,00
Gombar *et al.*[78]	95	SDA	electrotop., topol., MW, CMR	96,80	58	86,20	94,30
Cabrera *et al.*[83]	203	LDA	TOPS-MODE descriptors	80,50	40	77,50	81,82
De Cerqueira Lima *et al.*[84]	195	kNN	MolconnZ 4.05 Descr.	CCR_{train} 92,00	51	CCR_{test} 73,00	72,00
		BQSAR	Atom Pair Descr.	CCR_{train} 80,00	51	CCR_{test} 70,00	66,00
		SVM	VolSurf Descr.	CCR_{train} 88,00	51	CCR_{test} 81,00	78,00
		Bin DT	MOE Descr.	CCR_{train} 86,00	51	CCR_{test} 66,00	n.a.
Huang *et al.*[85]	203	SVM	Dragon calculated Descr.	80,80	40	90,00	91,00
		ML	2D and 3D descr.		40	80,00	82,00
Wang *et al.*[81]	206	SOM	Atom group, E-state	83,35	25%	n.a.	n.a.
				80,81			
	174	BPNN	Connectivity desc.	84,55	25%	n.a.	46,65
				75,31			29,01
Arodz *et al.*[82]	201	RFSBoost-LDA	as Xue	80,80	n.a	n.a.	n.a.
Zhang *et al.*[86]	171	DT	2D and 3D descr.	79,20	46	82,61	89,00
Yang *et al.*[87]	201	GA-CG SVM	as Xue	85,10			
Wang *et al.*[88]	212	SVM	global, size, shape, 2D Autocorr.	74,00	120	84,00	
		SVM	2D & 3D MOE	75,00		88,00	
		SVM	combination of Autocorr descr & MOE descr.	74,00		88,00	
		SVM	ECFP fingerprints	70,00		82,00	
Broccatelli[89]	150	NB	VolSurf descr. and 2D descr.	81,00	37	90,00	73,00
		kNN		83,00		78,00	
		SVM		81,00		78,00	
		RF		81,00		84,00	
		GA-kNN		83,00		81,00	
Poongavanam *et al.*[90]	282	kNN	functional group based fingerprints	73,00	202	67,00	74,00
		SVM		64,00		59,00	61,00
		RF		75,00		70,00	72,00

Table 1.2: Overview over classification studies performed. Taken and adapted from Demel *et al.*[45] Copyright Future Medicinal Chemistry. N - number of training set compounds, Acc. – accuracy, N (test) – number of test set compounds, ext - external, CV - cross-validated, SE – sensitivity, ADME – absorption, distribution, metabolism, excretion, toxicity; BPNN – back propagation neural network, BQSAR – binary quantitative structure-activity relationship, C-SAR – classification structure-activity relationship, DT – decision tree, accuracy and sensitivity values are given in percent, kNN – k – nearest neighbour, LDA – linear discriminant analysis, ML – maximum logarithmic, MOE – Molecular operating environment, PLS-DA: partial-least-square discriminant analysis, PNN – probabilistic neural network, SDA – spin discriminant analysis, SOM – self-organising map, SVM – support vector machine, TOPS-Mode : topological substructural molecular design, GA – genetic algorithm, RF – random forest, NB – naïve Bayes, RFS-Boost – random features subset Boosting, CG – conjugate gradient.

1.5 Similarity based approaches

The one aim of every researcher in drug discovery is ultimately the identification of new lead structures for drug development. For this purpose hundred thousands of compounds have been screened in high throughput assays and more recently in virtual screening. The one basic principle behind virtual screening is the similarity principle that assumes that compounds with similar properties share similar activities. This is a very tricky problem as Hugo Kubinyi portrayed[93] but nevertheless it represents the most promising approach to find hits and possible lead structures. An important aspect is that similarity can be defined in various ways like structural similarity or shape similarity and even the term similarity itself is entirely dependent on the similarity measure used, like Tanimoto or Tversky. The concept of shape similarity more and more moves into focus as new potential strategies emerge to use shape similarities in lead identification and optimisation. One of the most important aims in similarity based approaches is the potential of scaffold-hopping[94] and thereby finding potentially new active lead structures. An interesting article in this respect has been presented by Nicholls *et al.*[95] where the different aspects of shape in virtual screening, lead optimisation and protein crystallography are further elucidated.

Shape comparison methods can be divided into two groups, namely the alignment based methods and the moment based methods.[96] The alignment based methods rely on molecular overlay and encompass almost all the shape information but they do not contain shape in a numerical form. Moment based methods on the other hand contain shape information that is independent of any placement in space. They are easier and quicker to calculate but of course some information is lost on the way.

1.5.1 Alignment based methods

Jordi Mestres and colleagues proposed a molecular field-based similarity approach they termed MIMIC[97] which puts emphasis especially on an optimal alignment procedure of the molecules in question. The molecules are viewed

as Gaussian approximations. Molecular steric volume fields are calculated as shape descriptors and molecular electrostatic potential fields as chemical functions and suitably aligned. One molecule is held fixed while the other is moved over to find the best alignment position. After a gradient-seeking optimisation technique the respective similarities between the two molecules are regarded. The contribution of each of the molecular steric and electrostatic potential fields are weighted, customised and can further be inspected visually. In that way the contribution of each of the local similarity fields to the overall similarity between the two compounds is visible. It follows that structural aspects that contribute to the overlay of the molecules can be deduced and also those parts of the molecules identified that remain flexible and cannot be properly aligned. Those parts represent possible areas for future lead optimisation procedures. This concept has since been expanded to MIPSIM that takes into account the 3D representation of molecular electrostatic potentials inferred from either quantum chemical calculations and molecular interaction fields of a series of biomolecules.[98,99]

Over the years a number of shape based methods has been developed. One of them is the rapid overlay of chemical structures (ROCS)[100]developed by OpenEye which is based on Gaussian approximations and performs shape based screening of a query molecule over a database. Further information for this approach will be provided in the Methods Section 3.8.5 of this work.

Another commercially available program is Phase by Schrödinger.[101] It was developed for pharmacophore modelling but can be used for shape based screening of databases and also multiple binding modes can be detected herewith.[102] In a comparative study between Phase and Catalyst eight datasets were compared for 3 D QSAR generation.[103] In the study the unreliability of scoring functions was emphasised and in this respect the superiority of Phase over Catalyst acknowledged.

The method SHAFTS represents a hybrid approach for the calculation of 3D molecular similarity and was developed by Liu *et al.*[104] Shape Feature Similarity (SHAFTS) combines molecular shapes with pharmacophoric features, i.e. colored functional groups and utilises the feature triplet hashing

method from the PharmMapper Server for fast alignment. Optimal superposition is found between weighted volume overlap and the fit of feature points as Gaussian functions. A query molecule is determined with optimal conformation and pharmacophoric feature points calculated. After conformational sampling of the screened database pharmacophoric features are calculated and the respective triplets stored into a triplet hashing table. After searching the database for matching features an alignment routine is performed and the one alignment mode with the highest similarity score kept. The hybrid similarity score regards shape density overlap (ShapeScore) and weighted pharmacophore feature fit (FeatureScore). Though the overall approach and also performance of ROCS and SHAFTS is similar the difference lies in the alignment mode. ROCS optimises the volume overlap whereas SHAFTS enumerates all potential pharmacophore feature triplet matches. Interestingly when tested on a set of 2D fingerprints the 2D method in some ways outperformed the 3D shape method.

Another alignment based method recently was introduced by Hamza and co-authors[105] where the active ligands are ranked according to a scoring function and the structural diversity of the active ligands taken into consideration. They explore the pharmacological preferences of known active ligands and a weighted molecular shape density method is employed to determine the structural and chemical similarity. During alignment first the reduced structures are aligned which later are replaced by the full structures. Optimal superposition is determined via the self-developed HWZ function and a shape density model derived using atom-centred Gaussian functions. Good results could be obtained and prove less dependency on the structure of the query but regard common chemical information of the active ligands.

Other alignment-based methods include Shapelets,[106] flexible feature point pharmacophores[107] , fast 3D shape screening through alignment-recycling[108] and many more.

1.5.2 Moment based methods

Another idea for representation of molecules based on their topology was proposed recently by the Mestres research group.[109] Their principle aim was to skirt time and cost-consuming conformational sampling of molecules but in the same time utilise 3D similar information. They developed Shannon Entropy Descriptors (SHED) derived from atom-centred feature pairs not dependent on any conformation. In this approach four atom-centred features (hydrophobic, aromatic, acceptor and donor) are assigned, the shortest path length between two feature pairs calculated and the procedure repeated for a maximum path length of 20 bonds and for all possible 10 feature pairs. The information is stored in bins to generate a feature-pair distribution table and shannon entropy is calculated to depict the probable population of each bin. Used on compounds with different scaffolds but similar activities the depictive ability of the SHED descriptor could be proven. For all of those completely different molecules a similar descriptor profile resulted. Also in virtual screening the SHED descriptors could prevail with an enrichment factor of 0.8482. Pharmacological profiling that means predicting activities of known drugs on formerly disregarded targets is one of the new hot topics of in silico studies and also in this regard SHED profiles could show their merit[99] (Figure 1.8).

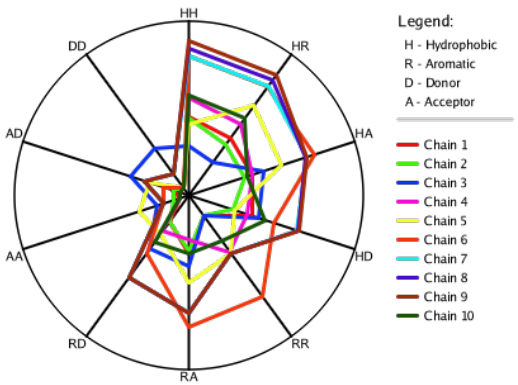


Figure 1.8: SHED Profile of 10 compounds of our in-house MDR-Database. Taken from Schwaha *et al.*[99] Copyright Scientia Pharmaceutica.

A completely different nevertheless clever approach is introduced with LINGO which is based on the SMILES string of molecules.[110] For this purpose the SMILES strings are divided into a number of substrings of specific length, the so-called LINGOs. Their ensemble and occurrences in the molecule depict the molecule specific LINGO. Using the LINGOs on the Tanimoto similarity measure the similarity of two molecules can be compared quite easily and also predictive QSAR models built (Q2 = 0.89 on logP prediction). Vidal *et al.* regard LINGO similar to a molecular fingerprint and propose respective use.[99]

In 2011 SheMS was presented to the community as shape based method using spherical harmonics (SH) for similarity comparisons.[111] Spherical harmonics are orthogonal spherical functions that can depict the shape of a molecular surface. Spherical Harmonics expansion coefficients are generated that can be used as descriptors and further be subjected to similarity comparison. Instead of using the coefficients directly for comparative purposes a weighted coefficient is derived enabling the user to specify the important areas. The respective weights are assigned depending on the conformation of the query and a set of reference conformations, active and inactive via a genetic algorithm. The resulting descriptors can be differentiated in lower band coefficients that represent the overall shape of the molecule and higher band coefficients that rely on specific features. By measuring the similarity between the SH shape descriptors a similarity score is derived. The method was tested on DUD sub datasets but as alignment-free method could not compete with alignment based methods like ROCS. Nevertheless it showed improvement over other non-aligned methods like the ultra-fast shape recognition algorithm and through PCA analysis could demonstrate enough shape information for shape comparison.

USRCAT represents an extension of the ultrafast shape recognition (USR) algorithm.[96] USR uses the topology of the bonded atoms to describe molecular shape. Hereby the distance between atoms and four reference points are calculated and finally a vector with 12 elements (USR moments) derived from the first three statistical moments of the distance distributions. This vector is unique for a set of coordinates. The CREDO interatomics database

contains the interatomic interactions between all molecules found in the 3D structures of the Protein Data Bank (PDB). Also all the ligands and residues in PDB structures are part of the CREDO database and it can be used to find ligands with similar activity profile for off-target effects. Up until then no tool was available allowing the user to search for ligands similar in shape in order to derive knowledge of biological targets or interactions. The ultrafast shape recognition with CREDO Atom Types (USRCAT) extension of the USR enables the user to find similar chemical entities in the PDB to a query compound. As the PDB houses many diverse ligands ranging from natural products to solvents simple similarity measures without regarding pharmacophoric features do not prove feasible. For this reason ultrafast shape recognition with atom types (UFSRAT) was developed that uses chemical features with atom types and ended in the final method USRCAT using CREDO atom types and user-specified weights for feature importance. This is a highly interesting approach and should be used in further studies. The PDB is always seen regarding the protein structures it contains but also many ligand structures in bioactive conformations can be found therein. Great chances for pharmacological profiling lie sleeping and special off-target effects can be studied.

The PubChem similar to the PDB represents a huge database full of chemical information that has to be made available for further use. PubChem 3D is available for neighbourhood searches like „Similar Conformers“ that allow identification of molecules with similar shape and pharmacophoric features and also a 2D tool using subgraph fingerprints is available.[112] The similar conformers tool employs the ShapeTanimoto measure which depicts the volume overlap of conformer A and conformer B. As this step is computationally very expensive for 28.9 million compounds in the database filter methods are used based on volume descriptors that dramatically reduce the cost. Molecular shape quadrupoles more or less depict the length, width and height of a molecule. These descriptors and simple volume descriptors were used for two conformers with the aim to install an additional dissimilarity filter to reduce computational effort even further. This could be achieved using a similarity threshold on volume information and molecular shape quadrupoles on

conformer pairs.

Other moment-based methods for similarity search include LASSO which uses surface point types to describe molecular properties and can further be used for QSAR studies.[113]

1.5.3 Comparative studies

Recently Distinto *et al.*[114] proposed the first purely ligand based virtual screening procedure combining shape-, 2D fingerprint and pharmacophore methods on the identification of HIV-1 reverse transcriptase dual inhibitors. After selection of a query compound for ROCS screening and conformational sampling of the database using OMEGA shape based screening was performed and the 200 best hits selected based on ComboScore. The four most active compounds found were further subjected to shape based screening, 2D ECFP fingerprint similarity searches and ligand based pharmacophore modelling using LigandScout. 27 compounds were selected by all three methods and finally a new scaffold could be identified. The identification of a new scaffold for HIV reverse transcriptase inhibitors, maybe binding to a different binding site of the protein, emphasises the high potential present in ligand-based, similarity based methods.

In 2007 Georgia McGaughey and colleagues[115] presented a study comparing topological, shape and docking methods in virtual screening and came to interesting results. Using the MDDR and the in-house Merck database 2D methods (like Daylight fingerprints, in-house TOPOSIM) to docking methods (FRED, Glide, in-house FLOG) and shape-based screening (like ROCS, in-house developed algorithm SQ) was performed on 11 protein targets. Their results emphasised that shape alone without any chemical typing does not nearly perform as well as the methods using pharmacophoric features. ROCS color outperformed SQW. Regarding the docking methods all of them were inferior to ligand based methods with FLOG performing poorest and Glide best. Also when regarding scaffold hopping no distinctive advantage for docking over ligand based methods could be detected when even 2D fingerprint

methods revealed different scaffolds. Interestingly the 2D methods could compete exceptionally well with the 3D shape based methods like ROCS color and SQW but though they were close, they could not present new compounds as structurally diverse as the 3D methods. Another interesting aspect is the low impact the source of conformer generation or the number of conformers actually has on the performance of the 3D ligand based methods. The race between ligand-based versus docking methods has clearly been decided for the ligand-based methods as docking generally performed poorer. This is especially important when considering the enormous time and computational cost necessary for docking when compared to fairly easily calculable ligand-based methods.

2

Aim of the study

ABCB1 has been studied for a long time by different groups and still much remains to be known. It represents a most challenging target for ligand based as well as structure based methods. In order to gain further insights into the substrate selection of this transmembrane transporter a number of ligand based classification studies is performed.

One of the limiting factors of each in silico prediction study is the quality of the data set used. Structures may be erroneously[116] depicted and often different labs use different assays under different conditions. A variety of differently assigned compounds is the consequence. In order to avoid this pitfall of literature compound selection the dataset used in this study is comprised entirely of natural compounds that have all been tested in the same laboratory under the same conditions. If mistakes have happened concerning these structures they should cancel each other out.

The most important aspect of this work is the similarity based approach used during all studies. The in-house developed SIBAR (Similarity Based SAR) descriptors have been proven a versatile tool in predicting ABCB1 inhibitors[117] and may enhance the probability of substrate prediction for so promiscuous a transporter as ABCB1. For this reason this work focuses on testing the limits of the SIBAR approach by using one diverse dataset. The underlying concept of SIBAR necessitates the selection of a suitable reference set beforehand, followed by the calculation of the chosen primary descriptor

types. The respective similarity values are then derived based on euclidean distances between compounds of the dataset and the reference set. These SIBAR values further are utilised as input variables for classification studies and machine learning. With SIBAR descriptors another window of opportunity is opened to adapt to the challenge at hand because additional to the non-specific classification method and the descriptor type a tailored reference set can be sampled from substrates and non-substrates of the respective dataset. The crucial steps remain the selection of the reference set and the method for calculating the similarities. One of the aspects of reference set selection has been shown recently on a set of propafenone inhibitors.[118]

There still exists a certain competition between the easy to calculate 2D methods and the computationally more intense 3D methods that may provide more structural diversity. The aim of this study among others was the exploration of 3D information used for SIBAR descriptor calculation and its impact on classification performance. Therefore, the performance of 3D versus 2D descriptors was compared and further a shape similarity method integrated into the SIBAR concept. The idea of shape similarity is highly interesting and promising as it uses just shape overlay and eventual pharmacophoric features to differentiate substrates from non-substrates.

In general the aim of this project can be summarized as follows:

1. Exploring the use of SIBAR based on a consistent data set.
2. Employment of shape similarity methods additional to SIBAR.
3. Establishing a valid classification model for ABCB1 substrate prediction.
4. Verifying the SIBAR approach on a set of very diverse natural products.

3

Methodological Background

3.1 Data set

As explained in the previous chapter many different assays for substrate/non-substrate/inhibitor determination are currently used. Whereas some assays provide the possibility to separate substrates from inhibitors (i.e. cellular monolayer efflux assay) others are not able to decisively determine substrates and inhibitors (i.e. ATPase activity assay). An additional problem also presents the application of these methods in different laboratories and also different protocols. These parameters make the accurate determination of substrates/non-substrates/inhibitors very challenging. Often compounds are labeled differently in literature changing from substrate to inhibitor and vice versa. Another aspect of the problem are the different nomenclatures used for ABCB1 ligands. Authors often differ in the correct label when using the umbrella terms substrate, inhibitor, non-substrate, modulator. In this study a substrate is pronounced substrate if transport by ABCB1 occurs. A non-substrate can passively diffuse into the membrane without any interaction with ABCB1. An inhibitor by its interaction with the transporter ABCB1 directly hinders transport of known substrates. A modulator or competitive substrate slows down transport rate considerably so that substrate compounds either are transported very slowly or not at all so that the passive diffusion rate is higher than the transport rate.

Therefore in order to avoid possible problems concerning substrates/inhibitor and non-substrate nomenclature the data set used in this work is based on the NCI60 screen published by Szakacs and colleagues.[119] In this work the authors took well known 60 human cancer cell lines (NCI-60) and tried to establish a link between ABC transporter expression and sensitivity to drugs. The goal was to determine which of the transporters do and which do not confer drug resistance or sensitivity to drugs. For this work all 48 known ABC Transporters with the exception of the 49^{th} transporter were screened and for measurement of the transcript expression rates quantitative real-time RT PCR (polymerase chain reaction after reverse transcription of RNA) employed. As ABCB1 is an export transporter inverse correlation between the pattern of transporter expression and the activity levels of its substrates is expected. Known substrates show negative correlation between substrate activity (paclitaxel, doxorubicin) and the level of transporter expression whereas known non-substrates (methotrexate, 5-fluorouracil) either are not correlated at all or positively correlated (Figure 3.1). A large data set consisting of 1429 compounds[120] was used for this purpose and in order to verify the approach the authors applied the MTT (3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolinum-bromide) assay on some of the top scoring samples. Correlation was computed with Pearson's correlation coefficients (48 genes x 1429 compounds) using bootstrap analysis with 10 000 iterations. The Pearson correlation coefficient provides a measure of the correlation between two variables and results in a value between +1 and -1. A positive coefficient shows direct correlation between the two variables whereas a negative coefficient depicts indirect correlation, the higher X the lower is Y.

For reasons stated above the data set of this work should be the result of one origin provided by one laboratory only. Therefore the compounds for the data set were taken from the work of Szakacs and colleagues.[119] Szakacs *et al.* defined a threshold of -0.25 to -0.3 for substrates and therefore compounds holding a correlation coefficient below -0.3 were assigned as substrates under the hypothesis: the lower the expression rate of ABCB1 the higher the toxic-

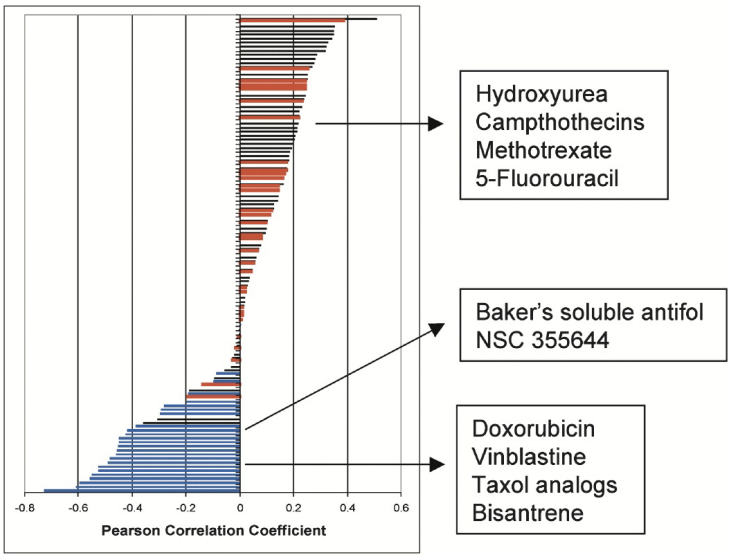


Figure 3.1: Taken from Szakacs *et al.*[119] Copyright Cancer Cell 2004. Relationship between drug sensitivity and ABCB1 expression in the NCI-60 for a set of 118 drugs of putatively known mechanism of action. Blue bars indicate known ABCB1 substrates, red bars indicate compounds shown in previous studies; not transported by ABCB1, black bars – no data known.

ity of the compound. Categorisation of non-substrates seemed to be a bigger problem as also positive correlation was observed by the authors resulting in higher toxicity in combination with higher ABCB1 expression rates. Consequently all compounds with coefficients between -0.02 and 0.02 were assigned as non-substrates following the meaning of the Pearson correlation coefficient as explained above.

Of course it has to be taken into account that the relationship between mRNA and protein expression may be influenced by some outside effects. Also cofactors like enzymes may show an erratic presence in the cells and noise could have been generated by calculating correlation over 60 different cancer cell lines. However, the natural product compounds of the data set were highly interesting and together with the aforementioned advantages we deemed the benefit higher than the risk.

In the end the data set used consisted of 240 compounds with quite a bal-

anced set of substrates and non-substrates considering the size of both classes. It is important to be careful in data set preparation as uneven distribution of classes may result in unsuitable bias of the model. Notwithstanding also the data set structures should be subjected to a curation process as Fourches *et al.*[116] have recently stressed.

3.2 Validation

The one aim of QSAR studies in general is to establish a reliable model in order to properly depict the relationship between the properties of a data set and its class assignments. Ideally the model is predictive enough to perform creditable virtual screening thereby decreasing the high throughput screening effort and cost. In order to be able to trust model predictions proper validation is of the utmost importance.

The workflow for QSAR model development as suggested by Tropsha[116] happens as follows. The first step is the curation of the original dataset and the splitting of the curated data set in a modelling set and an external validation set.[121] The modelling set is then further split into various training and test sets using leave-one-out (LOO) or leave-many-out (LMO) cross validation. In order to assess predictivity of the model against random assignments and for further use Y-randomisation is recommended. After that external validation on a set of compounds taken from the original data set and therefore undoubtedly within the applicability domain has to be done without fail. If the performance of the model is satisfactory it can be employed in virtual screening but bearing in mind the applicability domain of the model. After experimental testing of the compounds selected by virtual screening the model is good to be used in the pharmaceutical industry (Figure 3.2).

In order to judge a model's acceptability each research group has to define a threshold regarding the cross-validated and external validated accuracy of the model. As random accuracy is about 50% the accuracy threshold should be at least 60% in order to pronounce a model useful.

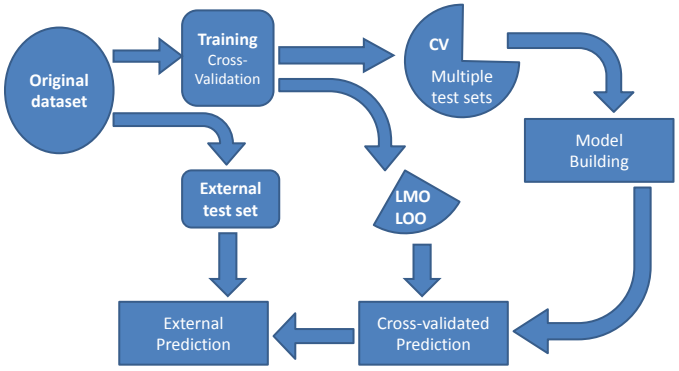


Figure 3.2: Workflow for validation: LMO – leave-many-out, LOO – leave-one-out, CV – cross-validation.

Validation encompasses the assessment of the model respective its robustness and naturally also its predictive ability. In former times internal validation alone seemed enough but experience has shown that internal validation alone is not a reliable enough parameter to illustrate a model's predictive power. Tropsha and colleagues emphasised the unconditional need for validation and cited in their study various examples for illustration.[122–125]

As explained above the compounds in this data set were either labeled substrates or non-substrates therefore no activity values were available and simple class labels assigned. For regression models the coefficient of determination r^2 explains the quality of the fit between the predicted activity and the measured activity of the training set compounds. For internal validation q^2 corresponds to the cross-validated coefficient of determination. The prediction parameters for classification were calculated based on the confusion matrix and consisted of the accuracy (A), the sensitivity (A1, accuracy on substrates), the specificity (A0, accuracy on non-substrates), the precision on substrates (Pr1) and the precision on non-substrates (Pr0) and the Matthews coefficient (MCC). These parameters are calculated using the number of

- true positives (TP): substrate compounds correctly classified as substrates,
- true negatives (TN): non-substrate compounds correctly classified,
- false positives (FP): non-substrate compounds classified as substrates and
- false negatives (FN): substrate-compounds classified as non-substrates.

Overall accuracy:

$$A = \frac{TP + TN}{TP + FP + TN + FN} \tag{3.1}$$

Accuracy on substrates (sensitivity):

$$A1 = \frac{TP}{TP + FN} \tag{3.2}$$

Accuracy on non-substrates (specificity):

$$A0 = \frac{TN}{TN + FP} \tag{3.3}$$

Precision on substrates:

$$Pr1 = \frac{TP}{TP + FP} \tag{3.4}$$

Precision on non-substrates:

$$Pr0 = \frac{TN}{TN + FN} \tag{3.5}$$

Also the Matthews correlation coefficient allows powerful assessment of predictive performance as it is a combination of precision and accuracy and regards also uneven distribution. Its values range from -1 to +1 where +1 depicts perfect prediction, 0 random prediction and -1 no prediction at all. A value over 0.4 is deemed predictive and the MCC is calculated as follows:[89]

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3.6}$$

3.2.1 Internal validation

Tropsha and colleagues[122] stressed the importance of validation. They divide the original dataset into an external test set and a modelling set which is in turn split into test and training set. The splitting of the modelling set in training and test set corresponds to internal validation or cross-validation. There exist a number of cross-validation combinations like leave-one-out (LOO) or leave-many-out (LMO) methods.

3.2.1.1 LOO and LMO

Leave-one-out means that all compounds of the modelling set minus one compound are taken as training set in order to establish a predictive model. Finally the compound formerly left out is used as a test compound in order to test the model's stability and predictive power. The whole process is repeated until every compound of the modelling set has been used as test compound. The modelling set in the leave-many-out method consists of all compounds minus a certain percentage (mostly 5 to 10%). Of course due to this procedure the training set for model development is diminished but stability and predictive power can be better evaluated with these methods.

3.2.1.2 Bootstrapping

Bootstrapping is another approach of internal validation. Bootstrap resampling tries to be representative of the population from which samples are drawn.[122] Therefore those samples are chosen at random with replacement. The major differences to LOO and LMO methods is the fact that not necessarily every compound is used as test compound during the iterations of cross-validation. Therefore some compounds can be part of the test set more than once whereas others never function as test substances at all.

3.2.1.3 Y-randomisation

This technique is widely used in order to test a model's robustness. First of all the activity vector is randomly scrambled several times so that the activities

annotated to the compounds in the data set are completely random. Then QSAR models are generated using this data set. Robustness is confirmed when low predictive statistics for cross-validated predictions and even test set predictions are rendered. Sometimes due to chance correlations or structural redundancy in the training set higher prediction statistics may occur but if all models obtained with Y-randomisation have relatively high predictive ability it suggests that no acceptable model can be obtained with the used method.

3.2.2 External validation

Tropsha and colleagues state that although low values of the cross-validated prediction statistics present a negative indicator for the prediction ability of the model high positive statistics do not necessarily promise high external prediction accuracies.[126]

As data sets with annotated activities mostly are scarce and the applicability domain of the model has to be regarded part of the original dataset is utilised as external validation set. Ideally a test set should contain about 15-20% of the entire dataset. Another bone of contention are the approaches used to generate the test set. Sometimes random selection through activity sampling and various clustering techniques is used whereas others employ completely different methods, for example self-organising maps. The criteria pronounced by Tropsha and colleagues respective the choice of training and test set consist of

- the chemical domain of training and test set must be close or identical
- the training set must be diverse
- the training set must contain at least 10 compounds of each class and the external test set at least five compounds of each class.

3.3 Bayes Theorem and binary QSAR

One of the old hats in classification methodology remains the Bayes theorem or naïve Bayes. It is often neglected as newer methods enter the scene but remains a reliable classification approach not only robust to noise but also quick to calculate.

3.3.1 Theory

The principle of the Bayes theorem is the calculation of probabilities P of class i. The prior probability where class i occurs in the molecules x is calculated by judging the information present in the training set. This information is then used following Bayes theorem in order to appraise the posterior probability of class i distribution.

Given a feature vector $X = (X1, X2, ..., Xn)$ of molecular descriptors independence of the features among themselves is assumed. The unobserved random variable C describes the class as either substrate (1) or non-substrate (0) and the probability *P(C|X)* can be given as

$$P(C|X) = \frac{P(X|C) * P(C)}{P(X)} \tag{3.7}$$

P(C|X) describes the conditional probability that the class hypothesis is true given example *X*. *P(C)* depicts the prior probability that the class hypothesis is right without any knowledge of the example *X*. In case of a multitude of molecular descriptors X represents their combination in total. The conditional probability of *P(X|C)* is the ratio of the total number of times X is observed as true to the class hypothesis over the total number of times the class hypothesis is true overall.[127]

X is predicted as substrate only if

$$f_b(X) = \frac{P(C = 1|X)}{P(C = 0|X)} \geq 1 \tag{3.8}$$

where $f_b(X)$ is called the Bayesian classifier.[128] Under the assumption that all attributes are independent of each other the classifier remains like

this:

$$f_{nb}(X) = \frac{P(C=1)}{P(C=0)} \prod_{(i=1)}^{n} \frac{P(X_i|C=1)}{P(X_i|C=0)} \tag{3.9}$$

with f_{nb} being the naïve bayesian classifier. Naïve Bayes provides the maximum of a posteriori probability hypothesis for the example x. Simply said:[129]

$$P(true) = \frac{likelihood(true)}{likelihood(true) + likelihood(false)} \tag{3.10}$$

$$P(false) = \frac{likelihood(false)}{likelihood(false) + likelihood(true)} \tag{3.11}$$

An important addition to the Bayes theorem is the zero-one-loss function which describes the error in terms of incorrect classifications. No penalty is given if the probability estimation is erroneous as long as the probability leads to the prediction of the right class of compounds. It follows that even if the probability estimation is poor the resulting classification could still be correct. This circumstance also explains why naïve Bayes is preferably used for classification studies and not for regression where its performance would not be as good.

The most crucial assumption as stated above is the fact that the features are independent to the value of the class variable. Obviously this proposition is very seldom true in real world conditions. That is the reason why the Bayesian classifier is also called Naïve Bayes or Stupid Bayes[130] and that may also be part of the reason why this method is not as popular as other methods in QSAR development.

Nevertheless Naïve Bayes has proven itself worthy of many a classification challenge in spite of inter-feature dependencies. Zhang and colleagues[128] have tried to gain some insight when the classification of Naïve Bayes is mostly affected by them. They suggested that Bayes could still operate satisfactorily if the dependencies are distributed evenly in classes or if the dependencies cancel each other out hereby making the distribution of dependencies the relevant factor. The authors also showed the optimality of naïve Bayes under

Gaussian distribution. Even distribution does not affect Bayes classification whereas uneven distribution may support a certain class. Inconsistent collaboration of the dependencies may result in their cancelling each other out. The study of Rish[131] supports this finding as they state that naïve Bayes overestimates the amount of information about the class that is present in the features. A reason may be that some information is counted twice due to the prerequisite of independent features. In their study naïve Bayes works best if the features really are independent of each other and also if they are functionally dependent. The performance weakens when between these two states.

Another study sang the same song. Pazzani and Domingos[132] observed that the bayesian classifier achieved higher accuracy than more sophisticated classifiers, for example C4.5 decision trees, in 16 out of 28 domains. In many of these datasets definite feature dependencies were present. However in spite of that naïve Bayes remained a very effective classifier.

Nevertheless problems[129] arise if a given descriptor value occurs only once in a data set and the class of the respective compounds is for example non-substrate. That means that for the classifier the probability of the test compound with that particular feature to be a substrate is zero. In order to glaze over this problem a Laplacian estimator is used. By adding a value of 1 to each probability $P(Xi|C)$ in the numerator and a value of N to the denominator, where N is the total number of pieces of evidence, a small but non-zero value emerges. Numerical values are handled by the assumption that they have a normal probability distribution.

Facts that undermine the positive aspects of a Bayesian classifier are:

- as an unsupervised learner no parameter tuning is necessary,
- robust to redundant variables, easy to use,
- no problem with missing variables and stable with noise present.[130]

3.3.2 Modifications of the Bayes theorem

There have been further developments that take into account the dependencies of the features among each other. One idea has been the augmented naïve Bayes where the class points directly to the attributes but there also exist links between the attributes themselves.[128]

Another approach are hierarchical bayesian networks. They build an extension of Bayesian networks and are able to work with structured domains. Through the introduction of a bias they seek to improve learning. Nodes in the hierarchical bayesian network represent aggregations of simpler nodes.[133] Other authors introduced the principle of rough sets working on decision tables and decision algorithms satisfying the Bayes' theorem.[134]

Bayes theorem has been fertile soil for automated classification systems like autoClass which was developed in 1988.[135] The authors present this system as unsupervised and also supervised classification algorithm working solely on Bayes theorem.

3.3.3 Binary QSAR

Making use of the advantages of Bayes theorem Paul Labute[136] developed the binary QSAR method which implemented the Bayes theorem in the Molecular Operating Environment (MOE)[137] of the Chemical Computing Group. He presented the new method as special feature suitable for virtual screening and as pre-selection tool of high-throughput-screening (HTS) candidates. HTS in the mean-time is associated with loss of precision because nowadays compounds only are categorised as either active or inactive. Therefore a binary classification method seemed to hit the mark. This method is based on the biased Bayes theorem that is defined even if a zero value should occur. The probabilities are estimated by accumulating a histogram of observed sample values on a set of B bins. This procedure is sensitive to the selection of bin boundaries since observations which occur close to the bin boundary are still treated as if they had occurred in the middle of one of the bins. Therefore a Gaussian smoothing parameter is introduced but even so some bins may happen to have zero values. The problem is solved by adding a constant to

each bin before normalising. During this procedure the class distributions for each descriptor are calculated with one distribution for active molecules and one distribution for inactive molecules in the training set (Figure 3.3).

```
Activity Field          : substrate
Smooth                  : 0.25
Condition Limit         : 1e+006
Component Limit         : 10

Active Observations     : 92
Inactive Observations   : 100
Observations            : 192
Descriptors             : 50
Components Used         : 10

                        QuaSAR Model      Chance
Total Accuracy          : 0.828125        0.503255
Significance (p-value)  : 2.191395e-019

Accuracy on Active      : 0.760870        0.421875
Accuracy on Inactive    : 0.890000        0.578125
Significance (p-value)  : 6.230039e-018


CROSS-VALIDATED STATISTICS
                        QuaSAR Model      Chance
X-Validated Total Accuracy      : 0.776042        0.503255
Significance (p-value)          : 4.035645e-014

X-Validated Accuracy on Active  : 0.706522        0.421875
X-Validated Accuracy on Inactive: 0.840000        0.578125
Significance (p-value)          : 1.124701e-012
```

Figure 3.3: Example of report file using binary QSAR – model of training set using SIBAR VSA descriptors.

The following computational procedure is observed:

1. each molecule is converted into a real vector followed by
2. a principal component analysis to produce the covariance matrix and then the identity matrix
3. the probability model is generated regarding Bayes theorem
4. estimating the class probabilities of a new set of molecules.

A similar approach of binary Quantitative Structure Activity Relationship has been implemented in Pipeline Pilot by Accelrys.[138]

3.3.4 Principal component analysis(PCA)

The ultimate goal of principal component analysis (PCA) is to find a way to re-express a data set by extracting the most meaningful contributions and dispose of the noise. The ultimate assumption in order to perform PCA is linearity of the data set. The data is shown as a linear combination of its basis vectors. The first function of PCA is to eliminate noise from the data set following the idea that the values with largest variances in the descriptor space express the truly interesting relationships. If noise is very high and no largest variance can be identified PCA is condemned to fail. However, if the existing variance is maximised the true dynamics of the field of interest can be visualised.

For this reason the basis of the matrix is rotated to lie in parallel to the best-fit line which in turn would give the true corresponding functionalities. The second function of PCA concerns the redundancy of the used descriptors. Highly intercorrelated descriptors are eliminated from the field leaving the most important features and thus visualising the true underlying dynamics. This is the true purpose of dimensional reduction.[139] Of the utmost importance in this context are the mean, the standard deviation and as a further deduction the variance.

The mean $\bar{x}$ is defined as the sum of all measurements x divided by the number n of measurements, giving

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{3.12}$$

The standard deviation is a measure of the margin of a data set. It is defined by

$$s = \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 (n-1)} \tag{3.13}$$

The variance represents another measure of margin of the data and is calculated as the standard deviation squared.

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{(n-1)} \tag{3.14}$$

The scope of this analysis is the examination of the relationship between various descriptors or measurements. The covariance is calculated as a measure of how much the dimensions vary from the mean.[140] It is a measure of the linear relationship between two dimensions. For a three-dimensional data set the covariance can be calculated between the x-y dimensions, the y-z and the x-z dimensions. The covariance consists of the summed up products of the variances between for example x and y dimensions divided by the number of measurements or descriptors as *n-1*.

$$cov(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1)} \tag{3.15}$$

Positive values of the covariance point to directly correlating data whereas negative values indicate indirectly correlated data. If the covariance is zero no relationship can be detected. The bigger the absolute value of covariance the bigger the redundancy. The covariance matrix is defined by its diagonal which represents the variance of the particular descriptors and by the off-diagonal terms characterising the covariance between the descriptors. Generally it is represented as an m x m matrix. The principal guidelines regarding the covariance matrix are as follows:

- large values in the diagonal terms hypothetically depict interesting relationships
- large magnitudes off the diagonal mean high redundancy

The optimal covariance therefore should describe zero values off the diagonal meaning no correlation between descriptors and each descriptive dimension herein should be ordered by rank respective their variance. In order to accomplish this, PCA assumes that all the vectors in the matrix are orthonormal. The main function of a PCA as described above is a rotation of the data to align its dimensions in a way of maximal variance.

1. The optimal normalised direction in the m-dimensional space is looked for along which the variance is maximised. This represents the first component.

2. As orthonormality is expected the second principal component is defined as the orthogonal direction from all previous directions along which variance can be maximised.

3. This procedure is repeated until all the vectors are selected.

The hypotheses on which PCA works are: linearity of data set, large variance corresponds to important relationships and the principal components are orthogonal. In order to derive the PCA *eigenvectors* and *eigenvalues* have to be regarded. Those two are always combined as there is no *eigenvector* without an *eigenvalue*. *Eigenvectors* encode the direction of the matrix and can only be found in square matrices. They are absolutely perpendicular to each other.

In order to perform a principal component analysis the following steps have to be performed

1. Subtract the mean from each of the data dimensions so that it represents the average across each dimension. Each of the dimensions now have an individual mean of zero.

2. Covariance matrix is calculated

3. *eigenvectors* and *eigenvalues* of covariance matrix are computed. They have to be presented as unit vectors.

4. The *eigenvector* with the highest *eigenvalue* is the principle component and all the other vectors are sorted according to their *eigenvalues* with highest to lowest.

5. Dimension reduction means not to include all the data of the original data set but to observe only the most important descriptors. In order to do that less principal components than previous *eigenvectors* are taken to be formed into a feature vector.

A drawback of PCA is the fact that if dimension reduction has been performed there is no way to return to the original data. Less dimensions cannot result in as many dimensions as before. Another disadvantage is

that if higher order dependencies are present than the elimination of second-order dependencies may not reveal the true relationships. Advantages of PCA are its simplicity to use with no parameters to tune and also its easy interpretation.

3.3.5 Applications of Bayes theorem

Many classification studies using Bayesian classifiers have been conducted. Here are some examples. Hongmao Sun[138] investigated the predictive power of a naïve Bayes classifier using Pipeline Pilot with in-house generated atom types. Using a set of 609 multi-drug resistance compounds atom types were assigned with the help of a classification tree resulting in 218 atom types from which 177 descriptors were derived. After defining a threshold for active and inactive labelling naïve Bayes classification was employed on a training set of 424 compounds. External validation with a test set of 185 compounds was performed rendering an overall accuracy of 82.2%. Also the enrichment curves were calculated proving the robustness and efficiency of the model in identifying the active or inactive compounds respectively.

Another study of Sun[141] was done on hERG compounds also using a Bayesian classification system. In this case 1979 compounds were used in the training set and atom types assigned via a classification tree. Again Naïve Bayes as implemented in Pipeline Pilot was taken and the whole model evaluated on an external test set of 66 compounds resulting in an overall external prediction accuracy of 87.9%.

In a study of Klon and collegues[129] three data sets (BBB – 129 training, 49 test set, Human Passive Intestinal Absorption Data Set – 205 training, 59 test set, Serum Protein Binding Data Set – 207 training, 53 test set) were taken and four different descriptor sets calculated. The software used to generate the Bayesian model were both Pipeline Pilot and MOE's binary QSAR and the in-house developed Bayesian classifier. Also a multivariate statistics approach was done by using the ADME Profiler. On the whole it was shown that the in-house developed Bayesian classifier performed best, seconded by the Pipeline Pilot and MOE's binary QSAR approach and lastly

the ADME Profiler.

Demichelis and colleagues[142] presented a study on classification of tissue microarrays to determine the relationship between tissue microarray and tumour heterogeneity. The study was carried out on the data of 35 and 34 patients and two proteins resulting in an accuracy of 65% and 58% respectively. Gao and colleagues[143] investigated the binary QSAR function of MOE on a training set of 410 estrogen analogues and an external test set of 53 compounds. They achieved an overall accuracy of 94%, an accuracy of actives of 78% and an accuracy of inactives of 98%. However, it has to be taken into account that only 62 actives versus 348 inactive compounds were present in the training set introducing bias into the system.

Paul Watson[144] presented a study also using an in-house developed naïve Bayes classifier together with in-house derived 2D Pharmacophore feature triplet vectors. These can be derived as follows. First of all different feature types have to be determined, like hydrogen bond donor features, ring atom features and so on. The next step comprises the calculation of the distances between all pairs of non-hydrogen atoms and in this manner the shortest possible bonded paths between non-hydrogen atom and atom-features are defined. These triplets of features are calculated and combined with the binned distances. Out of these a vector is generated where each value in the vector denotes the number of times a particular feature triplet occurs. The results have been promising and have been tested on two validation sets with the first one giving better external prediction results than the second one.

In 2008 Thai and Ecker[145] published a study of a binary QSAR model for classification of hERG channel blockers. In this study a set of 240 compounds were taken as training set with three different thresholds for activity. The descriptors used were two different kinds of van der Waals surface area descriptors as calculated with MOE. The best model of the first threshold achieved an overall accuracy of 89-94% on the test set of 58 compounds and the best model for the second threshold could accomplish 75% external overall accuracy though enhanced by a feature selection procedure.

These studies demonstrate the applicability of the Bayes theorem for modelling purposes and overall confirm the robustness and applicability of

the binary QSAR method implemented in MOE. On the whole one can say that binary QSAR indeed is a tool for rapid in silico screening purposes for real and virtual chemical libraries.

3.4 The Support Vector Machine

The Support Vector Machine (SVM) has been developed in 1995 by Vladimir Vapnik[146–148] and represents a supervised machine learning method. The purpose of an SVM is the supervised labelling of objects either to one class or to another. They can be used in regression and classification and are robust, reliable and able to deal with noisy data. Therefore they are widely applied in every possible field, i.e. text categorisation, hand-writing recognition, sound recognition, algorithmic training, in bioinformatics for predicting the suitability of proteins as drug targets,[149] for diagnostic purposes[150] and naturally also in ligand-based methods.

3.4.1 Theory

The primary principle behind the SVM is the structural risk minimisation principle where the generalisation error and the training error are held at a minimum. A hyperplane with maximum distance to the classes is built and projected into the multi-dimensional space in order to optimally separate the data in question. For prediction the test compounds are placed either to one side or the other of the hyperplane and the corresponding label is given.[151]

The probability of the test error is restricted by the frequency of the training error and the confidence interval (defined by the Vapnik-Chervonenkis dimension). The smaller the values of both the smaller also the probability P of error in the test set.

$$P(testerror) \leq frequency(trainingerror) + confidenceinterval \quad (3.16)$$

The Structural risk minimisation principle[152] is basically described by this

equation

$$R(\alpha) \leq R_{emp}(\alpha) + \sqrt{(\frac{h(log(\frac{2l}{h}) + 1) - log(\frac{\eta}{4})}{l})} \tag{3.17}$$

where $R(\alpha)$ represents the expected risk when learning to assign class y to the compound x. α represents the adjustable parameters. R_{emp} describes the empirical risk, the measured mean error rate of the training set. And variable l depicts the number of observations. The quantity

$$\frac{1}{2} \mid y_i - f(x_i, \alpha) \mid \tag{3.18}$$

is called the loss and in this particular case is situated between 0 and 1. Choosing some η fulfilling the conditions to be $0 \leq \eta \leq 1$ the probability for this case is $1 - \eta$ and the bound above holds.[146] Variable h is called the Vapnik-Chervonenkis (VC) dimension and comprises a measure of the notion of capacity. The VC dimension for the function $f(\alpha)$ is defined as the maximum number of training points able to be shattered by this function and therefore explains the hyperplane created in its course. If the VC dimension is h the principle holds true that at least one set of h points can be shattered but not necessarily every set of h points. The right hand side of Equation 3.17 pictures the „risk bound“. It is a bound on the risk taken and depicts a probability. The second term in Equation 3.17 on the right hand is called the „VC confidence“. Three essentials of this bound can be stated. First, it assumes that the training and test set are drawn from the same distribution of data but does not automatically assume normal distribution, second it is in most cases not possible to compute the left hand side and third knowing h makes it possible to compute the right hand side thereby presenting an upper bound to the expected risk. The VC confidence is a monotonic increasing function of h and holds true for any value of l. In order to diminish the actual risk as much as possible the right hand side of Equation 3.17, meaning the empirical risk and the VC dimension, should be at a minimum so that the upper bound of the actual risk is kept low.

The VC confidence naturally depends on the class of functions chosen

whereas the empirical and actual risk depend on the one particular function used in the training procedure. In structural risk minimisation (SRM) the goal is to divide the class of functions into subsets and further on to find the function minimal for the bound on the actual risk.[152] In general it can be said that the smaller the VC-dimension, the smaller the confidence interval but also the larger the value of error frequency.[148]

3.4.2 The separating hyperplane

In the multidimensional space of support vector machines a hyperplane is developed separating members of one class from members of the other class. In two-dimensional space this would be a simple line. The separating hyperplane chosen is the one with a maximum of distance between both sets of classes. In order to measure that margin some reference points from each class are needed. Those are the points nearest to the possible hyperplane and through them a hyperplane H_1 for one class and H_2 for the other can be drawn. These two form a parallel and no training points can be found in between. The training points building these two hyperplanes are called the support vectors (Figure 3.4).

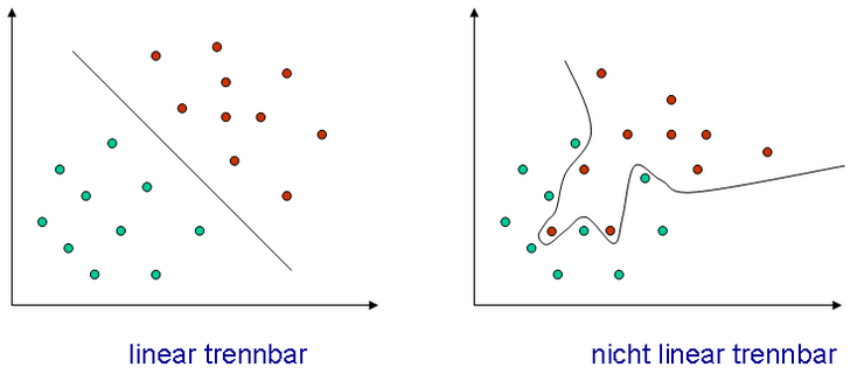


Figure 3.4: Separating hyperplane. Left – linearly separable, right – non-linearly separable. Taken from Wikipedia.[153]

Under the assumption that the hyperplane is able to separate the two classes of the training set without error an expectation value E can be cal-

culated that announces the probability Pr of error. The upper bound of this probability Pr is defined by the number of support vectors needed and their expectation value E and the number of training vectors in general, giving the following equation:[148]

$$E[Pr_{error}] \leq \frac{E[number \quad of \quad support \quad vectors]}{number \quad of \quad training \quad vectors} \tag{3.19}$$

The output of a linear SVM can be defined as

$$u = \vec{w} \cdot \vec{x} - b \tag{3.20}$$

where the variable w depicts the normal vector to the hyperplane, x are the input variables and the threshold b can be obtained from Lagrange multipliers. When $u = 0$ the separating hyperplane is derived and the support vectors lie on the planes $u = \pm 1$. The margin m is characterised through[146]

$$m = \frac{1}{2||w||^2}. \tag{3.21}$$

In linear cases the finding of such a separating hyperplane is very easy but the problem arises for datasets not linearly separable. In order to make that possible an important step is taken and the problem[154] is moved from a quadratic programming problem to a Lagrangian formulation. This bears the huge advantage that the training data will only appear in the form of dot products between vectors making projection in multi-dimensional space possible. The problem now is transformed into a dual and an optimisation problem with constraints.

However, more often than not the data analysed is non-separable instead of easily separable. In order to still find a solution and enable prediction with support vector machines a soft margin was developed. Non-separable data is characterised by outliers lying on the wrong side of the hyperplane. Nevertheless the support vector machine algorithm considering the whole data set performs well. The solution for this problem is a relaxation of the constraints and the introduction of a new cost, the regularisation parameter C. C is chosen by the user and the larger C the higher the penalty to errors.

3.4.3 The kernel trick

With some data it is impossible to find a separating line in two-dimensional space whereas in more dimensions this would be possible. To be able to project the data in multi-dimensional space a „kernel function“ or „kernel trick“ is needed. This is more or less a mathematical trick and can be performed with different kernels. Hope reigns that data non-separable in two dimensions may be separable in higher dimensional space without explicitly transforming the originally calculated descriptors (Figure 3.5). However this procedure also entails some dangers because projecting into very high-dimensional space can produce the so-called curse of dimensionality. This means that with the dimensions also the number of variables increases exponentially together with the number of possible solutions.[151] In consequence finding the right solution is a lot harder and may result in overfitting, i.e. only be applicable to the data set in question with no possible generalisation.

The most common kernels for these purposes[152] are the

- Gaussian radial basis function (RBF) kernel with width γ_i

$$K(x, y) = e^{-\|x-y\|^2/2\gamma^2} \tag{3.22}$$

- a linear kernel (polynomial kernel of degree 1),

$$K(x, y) = (x \cdot y + 1) \tag{3.23}$$

- polynomial kernel of degree ρ_i,

$$K(x, y) = (x \cdot y + 1)^{\rho} \tag{3.24}$$

- two-layer sigmoidal kernel.

$$K(x, y) = tanh(\kappa x \cdot y - \delta) \tag{3.25}$$

Noteworthy is also the fact that only through kernel functions data influences training and test functions.

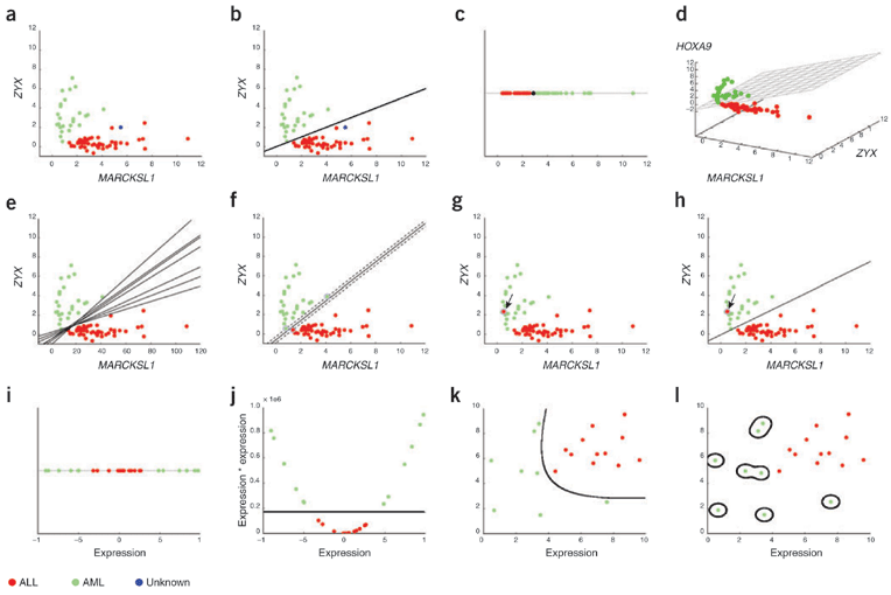


Figure 3.5: Taken from Noble[151]. Copyright Nature Comp. Biology. (a) Two-dimensional expression profiles of lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML) samples. Each dimension corresponds to the measured mRNA expression level of a given gene. The SVM's task is to assign a label to the gene expression profile labeled 'Unknown'. (b) A separating hyperplane. Based upon this hyperplane, the inferred label of the 'Unknown' expression profile is 'ALL'. (c) A hyperplane in one dimension. The hyperplane is shown as a single black point. (d) A hyperplane in three dimensions. (e) Many possible separating hyperplanes. (f) The maximum-margin hyperplane. The three support vectors are circled. (g) A data set containing one error, indicated by arrow. (h) A separating hyperplane with a soft margin. Error is indicated by arrow. (i) A non separable one-dimensional data set. (j) Separating previously non separable data. (k) A linearly non separable two-dimensional data set, which is linearly separable in four dimensions. (l) An SVM that has overfit a two-dimensional data set. In a, b, d–h, the expression values are divided by 1,000[151].

The challenge for the user concerning SVM consists of choosing the right kernel and finding the right values for the only two user-dependent parameters, namely regularisation parameter C and the one user-dependent variable regulating the kernel performance.

To find the ideal separating hyperplane is a difficult problem. In spite of the use of Lagrange multipliers the process involves quadratic programming and and renders a very time consuming quadratic programming problem (QP). With the Sequential Minimal Optimisation developed by John Platt the cost of time could been dramatically reduced.[154]

3.4.4 Sequential Minimal Optimisation

Originally Vapnik proposed so-called „chunking“ to solve this problem.[146,148] Chunking more or less breaks down the large problem in smaller bits with the purpose to identify all the non-zero Lagrange multipliers and also to remove all the zero Lagrange multipliers. At last all the non-zero Lagrange multipliers have been found and the QP-problem is considered solved. Though with this method the amount of data handled is significantly reduced, large-scale calculations still present a serious time-consuming problem.

In 1997 Osuna *et al.*[155] proposed a scheme for a new set of QP algorithms in breaking down one large problem into several sub-problems. If at least one violator concerning certain conditions (Karush-Kuhn-Tucker) is added to the examples of the previous sub-problem the overall objective functions will be reduced over each iteration and a feasible point will be held, placed within all of the constraints. The Karush-Kuhn-Tucker (KKT) conditions, taking into account some regularity conditions, have to be satisfied in order to find an optimal solution in non-linear programming.[156]

The Sequential Minimal Optimisation (SMO) algorithm takes into account Osuna's Theorem but SMO solves the smallest possible optimisation problem at every step. In this case the smallest possible optimisation problem consists of two Lagrange multipliers and at every step SMO chooses two Lagrange multipliers to optimise, finds optimal values for them and updates the SVM to the new values[154] (Figure 3.6)

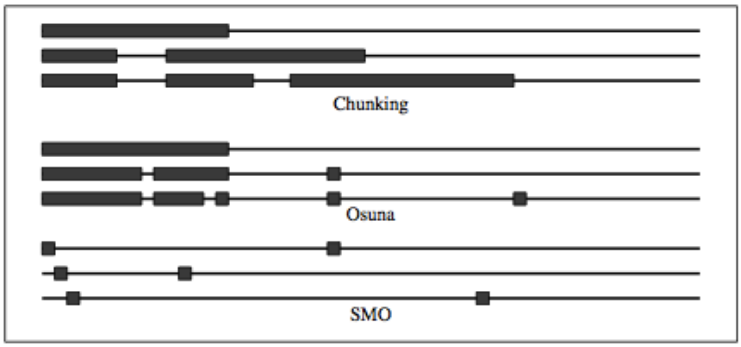


Figure 3.6: Taken from Platt.[154] Three alternative methods for training SVMs: Chunking, Osuna's algorithm and SMO. Three steps of every method are shown. The large boxes illustrate the Lagrange multipliers being optimised at each step. It can be seen that the number of Lagrange multipliers optimised with SMO is very small and therefore very fast.

This process reduces the quadratic programming problem to an analytically solvable problem and quadratic programming is neatly avoided. Platt states that even though more optimisation sub-problems are solved in the course of the algorithm each sub-problem is processed so fast that the overall QP problem is solved quickly. Also no matrix algorithms are needed and numerical precision problems are less likely to occur.

The Sequential Minimal Optimisation therefore offers an analytical method for optimising the two Lagrange multipliers and a heuristic method for choosing the respective multipliers.

The heuristic method for selection of the respective Lagrange multipliers is carried out as follows: The outer loop of the algorithm is concerned with the choice of the first heuristic and picks an example for optimisation that violates the Karush-Kuhn-Tucker (KKT) conditions. This will continue until all of the non-bound examples obey the KKT conditions within certain constraints ε between 0 and C. Bound examples are likely to stay put whereas non-bound examples are likely to move during the optimisation process. So the training set is thoroughly checked until all the subsets are self-consistent.

Platt compared his optimisation algorithm to Vapnik's standard chunking SVM learning algorithm and could show an impressing reduction of computing time. With a training set size of 1605 examples the SMO algorithm

was nearly 93 times faster than the chunking algorithm. Though SMO for non-linear SVM is only 15 times faster opposed to linear SVM where is is 1200 times faster than chunking SVM it represents an enormous progress in computing time. The time consuming task and the problem of quadratic programming presented a potential obstacle for the use of Support Vector Machines. The Sequential Minimal Optimisation of John Platt made the SVM approach available again for a number of classification challenges. One of the advantages of support vector machines is that only one minimum is found and premature convergence can be avoided.

3.4.5 Advantages and disadvantages

Support vector machine approaches comprise powerful tools for machine learning and are especially useful in non-separable datasets and in the presence of redundant features and noise. However, a distinct disadvantage of support vector machines is their black box function, making it hard for scientists to determine the relevant descriptors and consequently to interpret the models. The time needed to find the right parameters for optimal model performance also represents a drawback though nowadays easily available and fast methods can be found for parameter optimisation, i.e. grid.search function, etc.

3.4.6 Applications

An important influence of this work has been the publication of Bruce *et al.*[157] They strove to compare more easily interpretable ensemble decision tree methods including boosting, bagging and random forest, decision trees itself and the more difficult to interpret support vector machine. Eight data sets were used, consisting of angiotensin-converting enzyme inhibitors (ACE), acetyl-cholinesterase inhibitors (AchE), benzodiazepine receptor ligands (BZR), cyclooxygenase-2 inhibitors (COX2), a set of dihydrofolate reductase (DHFR) inhibitors, a set of glycogen phosphorylase b (GPB) inhibitors, a set of thermolysin (THER) inhibitors and a set of thrombin (THR) inhibitors. These were defined with a threshold as either active or inactive so

that the accuracy of the classification methods could be tested. The 2.5D descriptors have been used originally for the first data set. Also linear fragment descriptors containing data on atomic number, bond types, connectivity, chirality and number of hydrogen atoms were calculated. Working on the Java machine learning workbench Weka[158] they used the implemented support vector machine and also implemented J48 decision tree together with the ensemble methods boosting, bagging and random forest. The performances of each classifier was compared using 10-fold cross-validation averaged over 10 runs. Also the robustness of the models was determined with the standard deviation of the cross-validated accuracy.

The interesting part of this work was the assessment of the impact the changing of parameters can have on prediction performance. Rigorous multiple comparison statistical tests were employed. First of all the authors used two different kernels (radial basis function, polynomial kernel) for the SVM to see which kernel performed better in this instance. The two most important user dependent parameters as explained above are first of all the regularisation parameter (or complexity constant) C and the kernel dependent variable, for radial basis function kernel the width γ and for the polynomial kernel the exponent σ. For the regularisation parameter applies: the higher the value, the greater the importance of reducing misclassifications in the training model. The regularisation parameter was varied from the default 1 to 0.05 and 50. The exponent σ for the polynomial kernel from the default 1 was defined between 2 and 3. RBF width γ with the default 0.01 was chosen to lie between 0.001 and 0.1.

It was shown that non-linear SVM can enhance the performance for six of the eight data sets opposed to the default linear SVM implementation in Weka. Results for the tuned SVM in contrast to the default SVM has shown no consistent improvement in accuracy. In some instances the tuned SVM performed better and in others the basic SVM showed better results.

In the cases of random forest, boosting and bagging the number of trees in the aggregation is the user dependent parameter which was changed from

10 to 200 in steps of 10. The second alterable parameter, but only for random forest, are the number of features (descriptors) available during tree construction. In random forest only a random selection of descriptors are chosen to build a branching rule in a tree. M is the total number of descriptors. For the 2.5D descriptors this meant six or seven descriptors per node and those were increased to 10, 20, 30 and 40 descriptors 10 times for two forest sizes with 10 and 30 trees.

For the ensemble methods the increase of the tree size brought a significant increase of accuracy with a maximum reached between 30 and 50 trees. Additionally the robustness of the accuracies was improved by forest size with a definite improvement at 100 trees over that for 10 trees. Therefore an ensemble with 100 trees was considered ideal with a statistically significant increase in accuracy from 2 to 7% over the basic ensemble methods. Changing the number of descriptors for rule definition from default did not enhance the accuracy continuously and therefore the default setting was chosen.

Statistical tests have been performed like the two-tailed paired t-test, the Friedman statistic with a correction by Iman and Davenport and as a post-hoc test the Nemenyi test. The Friedman statistic with a correction by Iman and Davenport is able to state the difference between classifiers whereas the Nemenyi test should be able to detect which classifiers are significantly different from each other.

Overall it can be said that the support vector machine statistically performed best of all the classifiers particularly compared to a single decision tree though boosting, bagging and random forest often rendered similar and occasionally superior performances. The tuning of parameters did not bring a universally applicable set but confirmed that tuning was important for every individual data set. The interpretability of ensemble methods has to be taken with a grain of salt as there are differences between a single decision tree compared to a forest. The most important difference is due to „pruning“ that took place in the single decision tree and not in the „bushier“ random forest. An analysis of the frequency of descriptors chosen for the ensemble

methods did not bring the hoped result as no definite important descriptor set could be detected. A selection of descriptors is performed but those are not identical.

3.4.7 Software available

In the R-project[159,160] various packages provide the possibilities of a support vector machine[161] like e1071,[162] the kernlab package, the package klaR and the package svmpath. In this work the package e1071 was used which implements an interface to libsvm.[163] Libsvm represents a fast implementation of SVM formulations. It provides the most common kernels like linear, polynomial, RBF and sigmoid and the possibility of multi-class classifications as well as two-class predictions and regression. Its features include computation of decision values for prediction, class weighting in the classification mode, handling of sparse data and the computation of the training error using cross-validation.

The plot() method helps visualising data, support vectors and decision boundaries. The tuning of parameters is done using the tune() command which performs a grid search over the specified parameter ranges. The summary() command on the returned object lists the misclassification rate for each parameter combination and the best model. But tune() offers several alternatives (training, test set, leave-one-out error). An overall comparison of these packages yielded best marks for the e1071package and the kernlab package with a slight edge on the e1071 package.

Libsvm means library for support vector machines and was developed by Chih-Chung Chang and Chih-Jen Lin. In this version they implement an SMO-type Decomposition method.[164] The number of free support vectors is small and a technique called shrinking reduces the size of the problem without considering some bounded variables. If the shrinking process leads to errors the whole training set is re-optimised from the starting point of shrunk bounded variables. As this method is rather drastic it should be

applied only as long as necessary. After the stopping condition is reached caching is employed. It can reduce the computational time because instead of the full amount of data a temporary memory called cache is stored. In this algorithm a simple least-recent-use strategy is applied. Each structure corresponds to a kernel column and caches several elements of that column which results in different lengths of cached columns. In Libsvm two gradient reconstructions are done for training by the same method with the same amount of time. The best of the two methods is found and applied or at least a warning message indicates that the other method would be faster. Also libsvm enables parameter selection on the RBF kernel via cross validation and grid search.

Grid search is done as follows: The user specifies the parameter interval of the complexity constant and the RBF width within the grid space. Via trial and error of every grid point the best combination of the two parameters is identified via cross-validated accuracy. The user then performs training, cross-validation and external validation with the established parameters.

In WEKA[165] a support vector machine is provided together with a GridSearch function. The support vector machine makes use of Platt's sequential minimal optimization. Also it is possible to use a number of kernels as for example the polynomial kernel or the radial basis function kernel.

3.5 Random forest

3.5.1 Decision trees

3.5.1.1 Theory

Another method of classification are decision trees. They combine advantages such as easily interpretable rules with often unfavourable prediction accuracies. Decision trees start from a pool of training molecules and break them down at various decision points (nodes) into smaller portions. A node operates dependent on a certain descriptor and a certain threshold value, the

branching rule. When no further rule can be defined it is termed a terminal node or a „leaf“ and the molecules in this set are classified according to majority vote. These rules can be broken down to a certain classification rule and can help the user define certain properties an active or inactive molecule has to fulfil in order to be termed active or inactive.[157] This makes a decision tree attractive for every scientist as it can be made interpretable opposed to the mysterious black box the support vector machine presents. A major drawback of decision trees however is their classification performance which usually is considerably weaker than that of other machine learning systems like support vector machine. The classification and regression trees (CART) algorithm has been developed by Breiman *et al.*[166] and is capable of solving classification (categorical variables) as well as regression (continuous variables) problems resulting either in a classification or regression tree. Another common decision tree algorithm is C4.5.[167]

The first step of building a decision tree means the establishment of a deterministic value, the margin, for dividing one class of compounds (i.e. substrates) and the other class of compounds (i.e. non-substrates). This margin should be chosen in order to minimise the risk of falsely classified compounds (i.e. false positives, false negatives) as much as possible. In the following step all descriptors and all possible splits are evaluated and the descriptor best fulfilling these requirements is chosen as branching point or node. For these purposes different indices can be used. They consist of the Gini index, the twoing index and the information index as the three commonly used criteria.[168] The Gini index is a measure of distribution of classes. A zero Gini index means equal distribution whereas a Gini index 1 shows maximum unequal distribution which is preferable in a good classification node. For the number of available classes $j = 1, 2, 3, \ldots, k$ and the probability p of correct classification of class j at node t[169,170] the Gini index is defined as

$$\Delta i = 1 - \sum_{j=1}^{k} [p_j(t)]^2 \tag{3.26}$$

The information index is defined as

$$\Delta i = \sum_{j=1}^{k} -p_j(t) \quad ln \quad p_j(t) \tag{3.27}$$

The information index (also called Kulback-Leibler divergence) is a measure of the entropy gain and is equal to the total entropy of the descriptor if using it a classification can be achieved. In this case the relative entropies would be zero.[171]

The third step requires the actual dividing of the data set in two parts taking into account the calculated margin at this particular node. At each splitting point the group to be split is considered as mother group and the resulting branches as daughter groups. This procedure is repeated recursively for the resulting two branches excluding descriptors already used on the direct path to the present node whereas multiple use of descriptors leading to different branches or nodes is acceptable. The tree generation is terminated when either all the remaining compounds have been partitioned correctly or the number of available descriptors is at an end[172] (Figure 3.7)

The established tree is a maximum tree and therefore prone to overfitting. So-called pruning is necessary to remedy this circumstance and cut back excessive branches which ends in smaller sub-trees. In order to find the best sub-tree among them a comparison is done using the cost-complexity measure R_α for each tree which considers accuracy as well as complexity. R_α is comprised of the tree complexity (the total number of nodes of the sub-tree), the average within-node sum of squares and α, the punitive complexity parameter for each node. During pruning α is slowly increased from 0 to 1 and in this way a series of trees with decreasing complexity ascertained.[169] From these series the optimal tree is selected based on the cross-validated predictive error of these models. The predictive error consists of the overall misclassification rate for each of the subtrees. The optimal model is the one with a predictive error within one standard error (SE) of the minimal predictive error (one SE-rule). In this fashion a less complex model can be selected than the one with the least predictive error rate while maintaing reasonable

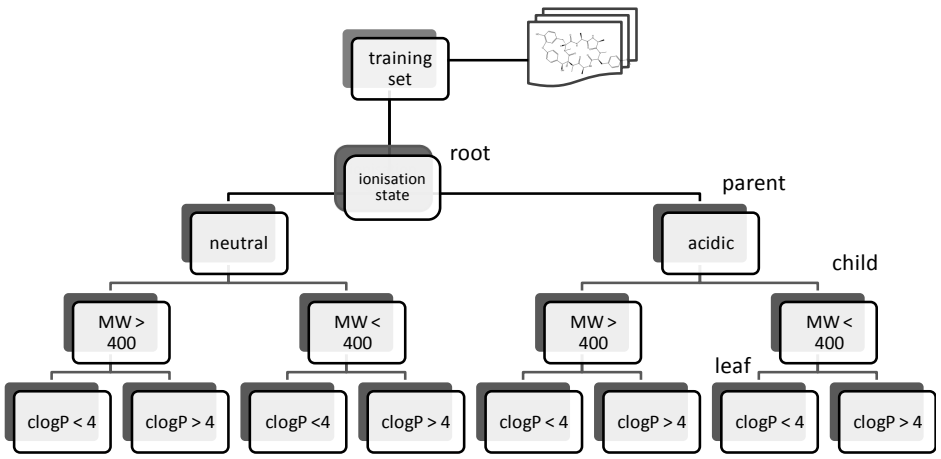


Figure 3.7: Schematic overview of a decision tree. Rules taken after ADMET rules of Gleeson.[54]

information and accuracy.

Excessive partitioning can also be avoided by appointing a user-defined maximum branching depth. The prediction accuracy can easily be calculated by a summation of the misclassified compounds in each leaf.

Classification with decision trees is a common occurrence as mentioned above they result in easily interpretable rules opposed to a black box like support vector machine, neural networks or others though a distinct disadvantage is the often bad prediction accuracy. As other possible disadvantages can be mentioned that the descriptor space is partitioned hyper-rectangularly and thereby a certain artificiality produced. Additionally the tree growing algorithm is greedy and may thereby overlook other solutions.

3.5.1.2 Applications

Andres *et al.*[172] have used decision tree models to predict the CNS permeability of drugs with an external predictive accuracy of 62.8% but interesting insights in classification determining descriptors. Another approach was done by Eric Deconinck and colleagues[169] who in an attempt to predict blood-brain barrier passage compared single trees with two different boosting approaches and achieved 84.4% prediction accuracy on a test set of 45 compounds with single decision trees. The best boosting approach received an external accuracy of 91.1%. Zhang and colleagues[173] compared single decision tree performance with a deterministic forest on the classification of gene expression data. Also in many other areas of research, for example in ecological prediction decision trees play a significant role.[174]

3.5.2 Ensemble methods

In order to strengthen the prediction accuracy of for example decision trees the use of ensemble techniques was proposed. The advantages of an ensemble technique only come to bear if there is discord in the classification votes of the individual classifiers. If one compound could not be assigned correctly by one classifier chances are that the other classifiers may produce correct

predictions and majority vote will rightly place the compound in question into the right class.

Three different reasons emphasise the idea of combined classifiers. According to Dietterich[175] first of all various hypotheses may be better able to find the „right“ hypothesis than one alone. Another reason is computational as many algorithms search for the local minimum but with more algorithms one may find the global minimum. The third reason regards representation of the true hypothesis. The idea is that a number of hypotheses may be better able to accurately picture the true hypothesis than one alone (Figure 3.8).

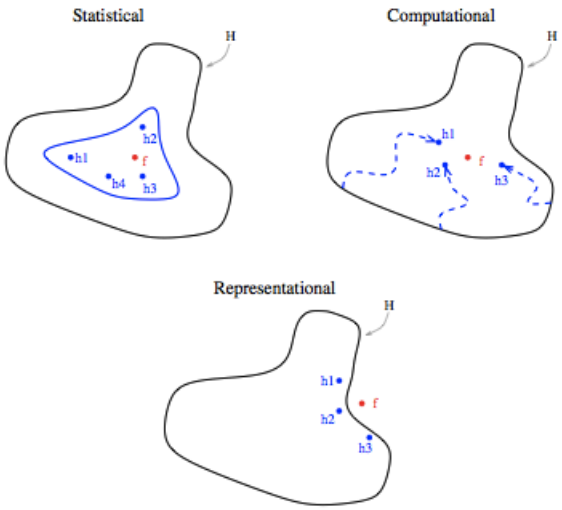


Figure 3.8: Taken from Dietterich.,[175] Copyright Springer Verlag Berlin Heidelberg 2000. Three reasons for ensemble methods.

Possibilities to build ensemble methods include the manipulation of training examples as can be done with boosting, bagging and even with x-fold cross-validation. An important prerequisite is instability of the algorithms used as otherwise their performance cannot be enhanced. Another way to increase performance is by manipulating the input features through feature selection. Randomness may also have a positive influence on performance as can be seen regarding bagging or random forest.

3.5.2.1 Bagging

Bagging was developed by Leo Breiman in 1996 and is an ensemble bootstrap method. The principle of bagging (bootstrap aggregating) is to generate multiple predictor versions and thus a plural vote for a given problem which should soften misclassification errors and enhance prediction accuracy. In bootstrap aggregating several „new“ training sets are generated out of the original training set by randomly drawing with replacement N examples out of it. This procedure results in a training set with the same number of compounds as the original training set but with some compounds being present multiple times whereas other examples may not be represented at all. Hence the ensemble consists of classifiers trained with different random versions of the original training set. The resulting training set is then used to train the classifier and the whole procedure is repeated several times (for example 50) thereby aggregated and then the results averaged (Figure3.9).

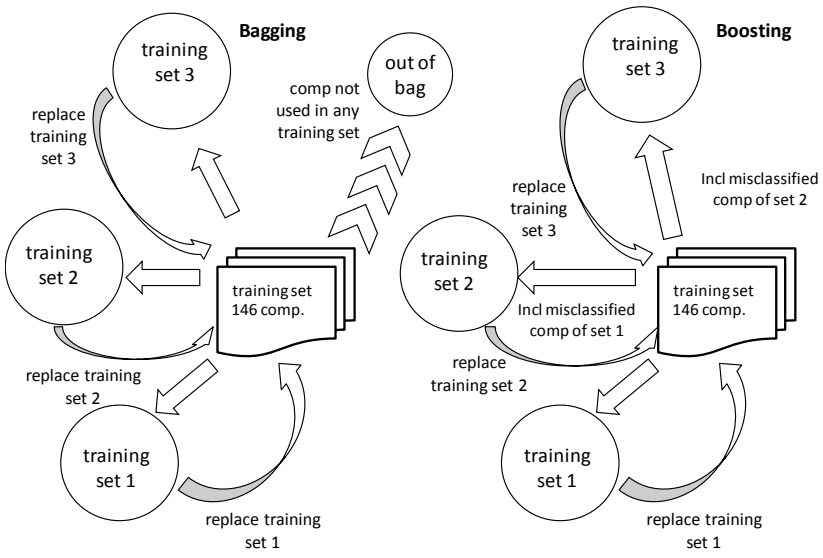


Figure 3.9: Schematic overview of the methods bagging (left) and boosting (right).

Another aspect of bagging represents the out-of-bag estimation, also proposed by Breiman.[176] The concept of bagging involves using some examples

of the original training set multiple times and others not at all. These "unnecessary" compounds could be used as an external test set on the classifier. Approximately two thirds of the original training set compounds are present in the resulting bootstrap training set whereas one third is not represented and therefore usable as external test set. For training set T with compounds x and class y a predictor Q is used. In bagging alternative training (1 to k) sets T_B are generated. That means that if $k = 100$ iterations are done approximately 33 of these compounds have not been used on the predictor $Q(x, T_B)$. These could be used as out-of-bag test sets and an overall out-of-bag estimation calculated. Out-of-bag estimates can be given for the generalisation error of bagged predictors.[176] With this estimation the user can test the classifier on a test set equally large as the training set. In classification procedures the out-of-bag estimates are unbiased and close to the prediction accuracy one would achieve with an external test set.

3.5.2.2 Boosting

Another approach opposed to bagging to enhance classifier prediction accuracies represents boosting. This method was developed by Freund and Schapire[177,178] . Adaptive Boosting (AdaBoost) takes a *weak* learning algorithm, that means performing slightly better than random, and through repetitive boosting enhances prediction accuracy. Similar to bagging boosting selects compounds of the original training set into a new training set but instead of random sampling this is done by weights placed on the training set compounds. These weights are assigned due to the performance of the training set compound on the former round of classifier prediction. Starting on equal weights in the first round compounds which have been wrongly assigned before are weighted differently in order to become part of the new training set so that the classifier has the chance to enhance its performance. This feature explains the name adaptive boosting which can transform a weak learning algorithm into a strong one (Figure 3.9).

Though AdaBoost seems a likely candidate for overfitting several studies showed that this is not the case. Instead it sometimes continues to lower

the generalisation error long after the training error had already converged to zero. This boosting algorithm performs a greedy search with respect to the misclassified examples. Advantages to AdaBoost are that there are no parameters to tune, it is easy to use and can be combined with more or less any method. A disadvantage of boosting is its susceptibility to noise as noise could send the algorithm to go round in circles.

Several comparisons between boosting and bagging have been done resulting in interesting conclusions. Richard Maclin and David Opitz[179] have compared bagging and boosting on 23 data sets and a neural network and decision tree C4.5. They found out that bagging achieves more accurate ensembles and is immune to overfitting. Boosting on the other hand showed sometimes a far better performance than bagging though its performance was varied and also susceptible to noise and despite everything the algorithm was supposed to overfit. Quinlan[180] did a similar experiment on boosting, bagging and C4.5 with a similar outcome. Both methods significantly enhance prediction accuracies when compared to a single classifier. Again bagging shows robust results whereas boosting's behaviour is erratic but occasionally significantly better than bagging. Leo Breiman in 2004[181] proposes that AdaBoost in its early stages gets closer to the Bayes risk.

3.5.2.3 Wagging

A method of combining boosting and a sub-form of bagging, so called wagging, has been proposed by Geoffrey Webb.[182] In this instance wagging assigns random weights to each training sets opposed to random bootstrap samples in bagging. He achieves greater mean error reductions than each of these methods alone.

3.5.2.4 Random forest

In 2001 Breiman proposed a new ensemble method called random forest.[183] This procedure is based on bagging but with a further enhancement. Additional to bagging that means random sampling out of the training set with

replacement, also the best descriptor used for the split is chosen from a random set of descriptor variables (*mtry*) instead of all the available descriptor variables (p). A defined number of trees are grown to a maximum employing CART algorithm until no further split is possible and no pruning takes place. Class prediction is done with majority vote in classification and by averaging their outputs for regression purposes. The margin function in this instance describes the average ratio of right class prediction of the classifiers versus wrong class prediction of classifiers. This means the function also shows the confidence of the classification and describes the correlation of the classifiers. As a result the upper bound for the generalisation error (PE) is determined by the margin function, i.e. their correlation r, and the strength of the classifiers s used and works similarly to the VC-type bounds of for example the support vector machine.

$$PE \leq r(1 - s^2)/s^2 \tag{3.28}$$

Pointing to the Strong Law of Large Numbers Breiman absolutely negates an eventual tendency to overfit in random forests regardless of the number of trees used. In his paper Breiman distinguishes two different random forest variations: Forest-RI and Forest-RC. In Forest-RI (random input selection) the simplest kind is produced by randomly selecting, at each node, the group of variables for a user-defined group size (*mtry*) used for the split. Forest-RC (random linear combinations) works by defining the number of variables to be combined (L) which at a given node are added together with coefficients that are uniform random numbers on [-1,1]. The number of combinations again is defined with (*mtry*) and the best combination is chosen to be employed for the split. L and *mtry* can be user-defined though Breiman's studies result in a relative insensitivity of the generalisation error to any change of *mtry*. Consequently the size of the random variable subset is fixed, *mtry*, with a default value of $p^{1/2}$ for classification and $p/3$ for regression.[184] The variable p depicts the number of overall available descriptors (Figure 3.10).

If categorical variables are present in the input selection they have to be made comparable to numerical variables. This is done by selecting a random

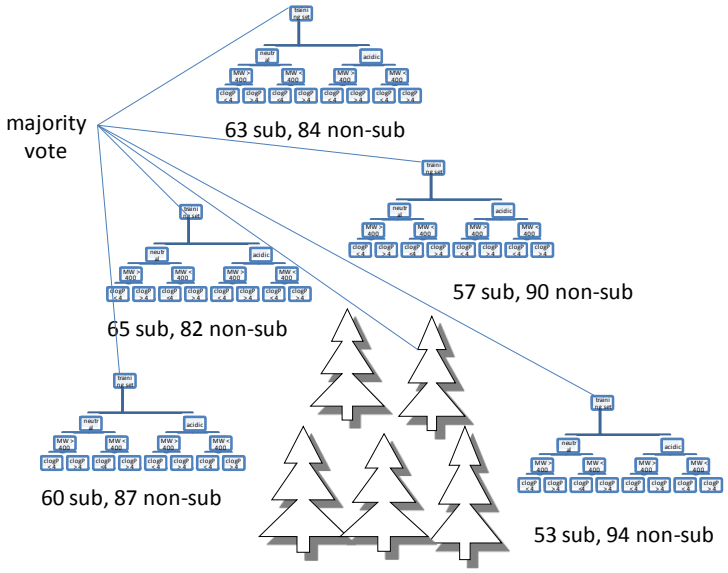


Figure 3.10: Schematic representation of a random forest.

subset of the categories of the variable. Then define a substitute variable that is one when the categorical value of the variable is in the subset and zero outside.[183] A consequence is that in this case number of trees F must be considerably increased in order to provide good external prediction accuracy.

The advantages of random forest regarding especially AdaBoost are as follows: Accuracy (at least as good as AdaBoost), robustness and speed (also compared to Adaboost which is slower owing to the determination of the to-be-assigned weights). Breiman goes as far as to propose that AdaBoost in reality functions as a random forest where the weights on the training set are selected at random from its distribution. The difference being in the fact that random forest really is random selection of variables whereas Adaboost's selection of weights will be coupled to the training set.

Variable Importance

It further was proven that random forest is more robust to noise than Adaboost.[183] The employment of bagging in random forest puts to use its advantages like further increase of prediction accuracy and the unbiased out-of-bag estimates. These provide an approximation of the generalisation error and may also be used to derive the variable importance. The measure of variable importance is generated by permuting each of the values of the descriptor variables in turn for the out-of-bag data. Thereby any existing correlation between variable and to-be-predicted class is considered gone. Then the data set is run through the classifier and the classification decisions saved. This is repeated for every descriptor variable and the out-of-bag class votes for each changed variable are compared to the real class label. Out of this the misclassification rate can be calculated and the increase for each changed variable viewed. The variable with the most impact on the classification rate is top in the list of important variables. Another feature that can be calculated with random forests are proximities. They originally formed a NxN matrix and are derived by putting all the data including training and out-of-bag data down the tree. The proximity between two molecules is the number of trees they land together in the same node over the number of trees in general and calculate in this way the intrinsic proximity. This can also be done with an external test set.[185]

Bias and variance

Important words to measure classifier performance are bias and variance. In the classical sense bias means the preference of one hypothesis over another without cause by observing the training data set. They can be defined as *absolute* and *relative* biases. According to Dietterich[186] in an absolute bias the learning algorithm makes the assumption that the function to be learned belongs to a marked class of functions and thereby eliminates all other possibilities. The relative bias on the other hand is the assumption that the to-be-learned function is more likely to be part of one set of functions and not another. He states that decision tree algorithms in learning (CART, C4.5) prefer smaller trees to larger ones and if a small tree suffices a large

tree is not even tried. This bias is also referred to as the machine learning bias as its functions may be desired by the user of the algorithm.

Another kind of bias is represented by the statistical bias which describes a systematic error of the learning algorithm. The variance on the other hand focuses on the variation in the algorithm due to different kinds of training sets, noise or a random behaviour of the learning algorithm itself.

$$\text{Error} = \text{Noise} + \text{Bias} + \text{Variance}$$

It follows that the smaller the variance and the smaller the bias the smaller also the classification error. If for example the drug space can be depicted better the algorithm could reduce its statistical bias. Whereas if the alternative results for an algorithm increase or the available training set samples diminish then the variance of the algorithm increases. As the removal or addition of one single training compound may influence the choice of split criterion in decision trees this is cause for a great variance. Pruning was thought to cure the problem as high risk leaves with high variance could be plucked. Interestingly in the study of Dietterich pruning did not have a sinking effect on the variance but on the contrary increased the bias and in the long run does not improve classification accuracy. Bagging interestingly slightly increased the bias but reduced the variance nearly by half thereby rendering the classification prediction significantly improved. Boosting is able to decrease both variance and bias but bias reduction is greater than variance reduction.[182] Single trees are thought to be low-bias but with high variance. With a forest however the variance can be decreased by averaging over the ensemble of trees while the bias is kept low.

Performance

In three different studies Vladimir Svetnik[184,187,188] and colleagues compared the aforementioned approaches with other QSAR methods like Recursive Partitioning, Partial Least Squares and single Decision Trees. An advantage of random forest and also decision trees is the fact that a large number of

descriptor variables, including a significant amount of irrelevant descriptor variables, can be used without prior descriptor selection whereas in linear discriminant analysis (LDA) and k-nearest neighbour preselection is sometimes needed if the number of irrelevant descriptors is significant.

In a comparative study of random forest together with recursive partitioning and partial least squares six different available data sets were used (blood brain barrier permeability, estrogen receptor, P-gp, MDRR and COX-2 inhibition) with five-fold cross-validation. As shown previously[183] and confirmed again by Svetnik and colleagues[184,187] random forest mostly shows no significant sensitivity towards the size of the random descriptor group important for splitting nodes (Figure 3.11). But if many irrelevant descriptors are present this could be important after all. Other parameters to tune in random forest include the number of trees and the minimum node size. In order to get stable results Svetnik proposed 1000 trees adequate for the P-gp set if the goal is accuracy rates but 10000 trees if the proportion of votes for a class is the subject of interest.

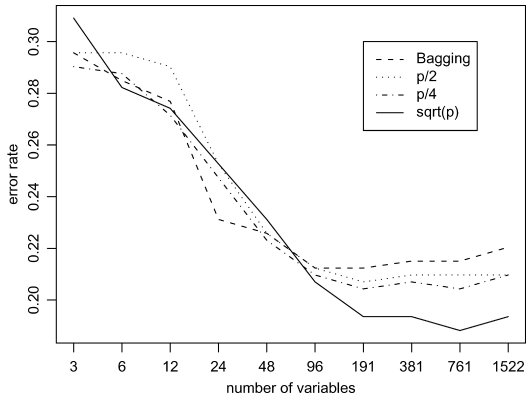


Figure 3.11: Taken from Svetnik *et al.*[184] , Copyright J. Chem. Inf. Comput Sci 2003. This figure depicts the median cross-validation test error when halving the number of important descriptors via *mtry* parameter based on the ABCB1 data set. The default *mtry* parameter $p^1/2$ seems to perform best as reduction up to 191 descriptors is possible without enhancing the error rate.

Another changeable parameter represents the minimum size of nodes which by default is one in classification and five in regression. Below this

threshold no split will be possible. Using the variable importance ranking as feature selection procedure did not enhance prediction accuracy and calculated on the out-of-bag error leads to severe overfitting. Trials to tune the forest by changing the default of *mtry* from $p^1/2$ did not lead to better results but performed slightly worse. The study of Svetnik and colleagues concluded that random forest is safe to use off the shelf without any parameter tuning necessary and still achieves excellent prediction accuracies compared to other methods.

Random forest calculation is also very fast because as only *mtry* numbers of descriptors are used for splitting the computational cost can be lowered especially as no pruning is done. As a result random forest occasionally can be calculated faster than a single decision tree.[184] In a direct comparison between boosting and random forest also other approaches like support vector machine, decision tree, k-nearest neighbours, naïve Bayes, support vector machines and partial least squares were examined[188] with 10 data sets. As expected boosting and random forest demonstrated the best prediction accuracies on the whole but showed a definite disparity. The study showed that for smaller data sets and classification purposes random forest seems to be more appropriate than boosting. Additionally boosting has more parameters to be optimised which could be time consuming. Second in QSAR methods came the support vector machine with the radial distribution factor kernel followed by the linear kernel support vector machine.

Disadvantages

The opportunity to compute the variable importance of individual descriptors has been put under scrutiny by Strobl and colleagues.[189] In their work they illustrate that the variable importance measures based on Breiman's original CART algorithm are not reliable if the predictor variables differ in their number of categories or their scale of measurement. As mentioned previously three kinds of measures of variable importance exist. The first is simple counting how often a variable is selected by the trees in the ensemble,

the second makes use of the Gini criterion and the third is the permutation accuracy importance measure that has been described in 3.5.1.1. The Gini index is known to prefer variables with a higher level of categories to others with less categories.

Strobl *et al.* propose a different method for calculation of variable importance called cforest. Hereby the split criterion is not the Gini index as with CART algorithm but a conditional inference framework called cforest. Two simulation cases were built. The first case was built with variables absolutely not correlated with the class to be predicted. In this instance the generation of variable importance measures should not be possible (Figure 3.12). In the second case only one descriptor variable was correlated with the predicted value. A reliable variable importance measure should be able to pick the right descriptor variable. The results disappoint as all the variable importance measures fail but none as much as the Gini index criterion followed by simple counting and last Breiman's permutation importance measure. In the first case where no variable importance should have been detected a clear bias regarding variables with more categories could be observed which was visible in all three detection criteria. The scaled variable importance is given by default (*z-score*) but the authors advise not to interpret the magnitude of the scaled variable importance as it depends on the number of trees in the forest. The second case study again did not provide the right results straight away. The random forest permutation importance function could identify the relevant descriptor but the preference of variables with more categories could strongly bias the results. Scaling seems to lessen the problem a bit. Their own approach of cforest performs better but only when not bagging is used but sampling of the training set without replacement.

With these experiments another danger of random forest could be disclosed. If the variable importance prefers variables with more categories then also the forest itself may place more importance on variables with more categories than others and thereby falsify the prediction performance. A chance to circumvent this phenomenon may be that as random seeds are used to produce random forest variable importance measures are only reliable if the method is repeated several times with different random seeds.

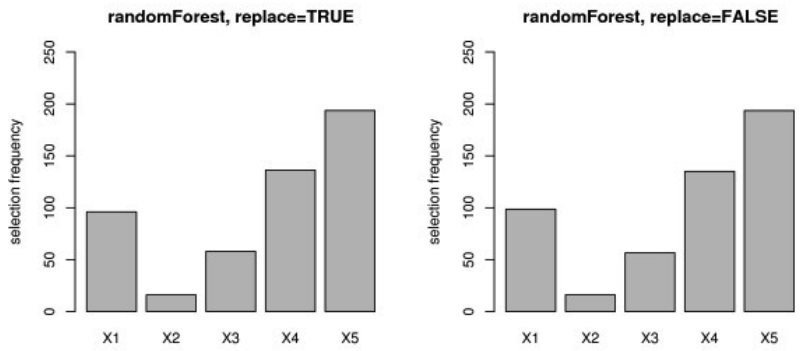


Figure 3.12: Taken from Strobl *et al.*,[189] Copyright BMC Bioinformatics 2007. Results of the null case study – variable selection frequency. None of the variables should be deemed informative. The figure depicts the frequencies when random forest function is used. The left hand side displays bootstrap sampling with replacement and the right hand side illustrates subsampling without replacement.

The bootstrap sampling with replacement itself as shown in the study seems to enhance the bias towards variables with more categories. In conclusion the published results show that if variables with the same number of categories are employed the variable importance measures are not affected by bias. If this is not the case then the random forest variable importance measures can be unreliable and even misleading. In another work of the authors[190,191] they explore the eventual bias if variables are inter-correlated. A spurious correlation may occur when a correlated descriptor variable by its correlation with another really influential descriptor variable appears more important than it is because the original influential descriptor is not present in the model.

This is a dangerously misleading development and especially gains in importance when *mtry* is small and chances are high that the original influential descriptor is not present in the subset from which to choose. With the increase of *mtry* chances rise that descriptors correlated with the highly influential descriptors are chosen more often than others that are uncorrelated but may be higher influential. To remedy that fact they propose a scheme of

conditional permutation where the variable x is permuted within groups of z where z can contain large sets of covariates of different scales of measurement. Hereby a grid should be made available that is better applicable for different types, more penurious and computationally effective.

Applications

Forests have been used with success in a number of studies. Tong and colleagues[192] employed a decision forest on an estrogen data set containing 232 compounds and calculated 202 descriptors with Cerius 2 software. The decision forest is more or less a boosted forest as for each tree a distinct set of descriptors was used which was subsequently excluded from use for all the following models. The quality of the models received should be comparable so that each model can significantly take part in the final prediction. After a tree is built boosting was used in order to get the best prediction performance. Another distinction to random forest is the pruning that takes place further on for each tree. Their approach showed significant improvement when compared to a single tree with a distinctive reduction of the misclassification rate.

Random forests have also been used with natural products as in the study of Ehrman and colleagues.[193] In this work a database of 240 chinese herbs containing 8264 compounds has been screened with random forest on 10 respective targets with most satisfactory results. Literature search detected evidence for 83 herb-target predictions.

DeLisle and colleagues[194] have proposed the induction of decision trees with evolutionary programming contrary to the traditional recursive partitioning. They argue that recursive partitioning is greedy in selecting splitting variables which could be altered by evolutionary programming. This was shown on two data sets (300 compounds and 436 compounds) and 25 descriptors previously selected using simulating annealing.

Other approaches on feature selection are presented in the work of Blum and colleagues.[195] Their approach of evolutionary programming consisted at first of the generation of random decision trees with the training set. Mild

restraints concerning the minimum number of compounds in nodes or leaves are given. After that a fitness score selects two members of the forest which are then copied and returned to the original forest. The acquired copies are then mutated or crossed over. Crossover means that one part of a tree is swapped with another part of the tree. Mutation changes the values of the descriptors at nodes. The result are two new decision trees and the procedure is repeated until the number of children trees equals the number of the original forest. After that the original forest is deleted and replaced by the children forest. This is continued as desired. Results showed a significant reduction in complexity of the trees though the prediction accuracies were comparable with other tree ensemble methods.

The package randomForest[196] was implemented in R[159] providing an interface to the Fortran Code by Breiman and Cutler.[185] In this function the calculation of variable importance is possible as well as proximity measures and their paper also makes available some tips for practical use.

3.6 Other classification methods

3.6.1 Linear discriminant analysis

Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are two similar techniques that are used for different purposes: reduction of dimensionality and classification. LDA is able to process data with unequal distribution of classes. Hereby the ratio between classes is maximised and separability enhanced. A typical application for linear discriminant analysis is speech recognition. LDA and PCA are characterised by different aims. PCA is primarily used for feature classification and removes redundant information whereas LDA's first priority is data classification. Using PCA the data is transformed and consequently the location of the data is changed. LDA on the other hand does not switch the location of the data but tries to draw a decision-region between the two or more classes at hand.[197] It should provide better understanding of the distribution of one's data. LDA assumes

a multivariate normal[198] between each group and that each group shares the same covariance matrix.

Comparable to binary QSAR this method also works on basis of the Bayesian principle. First of all the mean of each data set is calculated together with the mean of the overall data set and prior probabilities calculated. After that criteria for the separability of classes have to be found by evaluating the compounds placed between classes and within-class. The covariance of the respective class is the within-class scatter. It is calculated as follows:

The covariance matrices are considered symmetric.

$$S_w = \sum_j p_j \times (cov_j) \tag{3.29}$$

and for two classes the covariance matrix looks like this:

$$S_w = 0.5 \times cov_1 + 0.5 \times cov_2 \tag{3.30}$$

The objective of LDA is to maximise the distance between class scatter to the within class scatter. The resulting solution for this problem characterises the axes of the transformed space. The scatter S that lies between classes is calculated like this:

$$S_b = \sum_j (\mu_j - \mu_3) \times (\mu_j - \mu_3)^T \tag{3.31}$$

If LDA is of class dependent type then for each of the classes separate optimising strategies are required which can be derived by

$$criterion_i = inv(cov_i) \times S_b \tag{3.32}$$

A set of *eigenvectors* with non-zero *eigenvalues* are linearly independent and therefore stable in transformation. By using linear combinations of the *eigenvectors* any vector space can be illustrated. If linear dependency is

present then *eigenvalues* of zero are the consequence. They have to be removed in order to guarantee non-redundant features.

For any *L*-class problem *L*-1 non-zero *eigenvalues* are produced. For a two-class problem one significant eigenvector is found where maximum discrimination is possible (Figure 3.13). By transformation of the data to one axis definite distances between the classes can be detected. Transformation proves to be especially important when the classes are overlapping and no significant distance between classes can be observed.

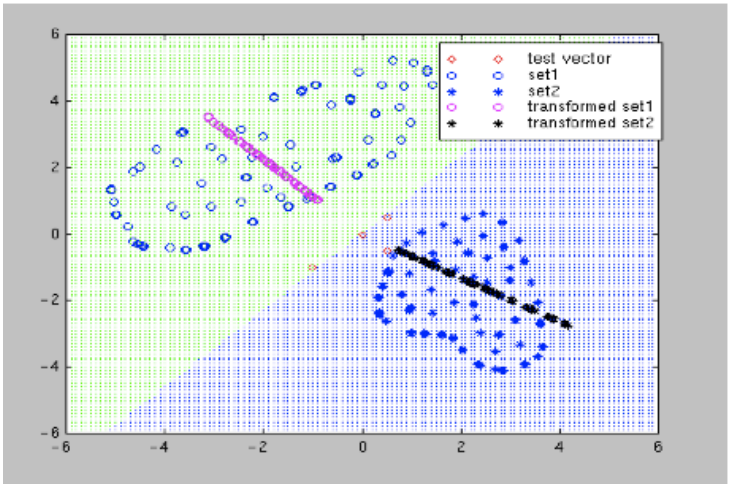


Figure 3.13: Taken from Balakrishnama *et al.*[197] Data sets illustrated in original space and also transformed space along the transformation axis for class-dependent LDA of a 2-class problem.

After LDA transformation euclidean distances or root-mean-square distances between data points are calculated between the mean of the transformed data set and the test vector. The smallest euclidean distance among the classes defines the characteristics of the test vector. It can be assigned as either class α or class β, rendering posterior probabilities. In LDA either class-independent or class-dependent transformation is possible and selection is depending on the aim of the study. For generalisation purposes the class-independent type is preferred whereas the class-dependent transformation guarantees better discrimination between the respective classes.[197]

A recent study showed that sometimes LDA is inferior to other machine learning techniques like random forest and in that special case did not perform much better than random.[199]

3.6.2 Quadratic discriminant analysis

The performance of linear and quadratic discriminant analysis depends on the size of the training set and ultimately whether the distribution of classes follows a multivariate normal distribution. The discriminating parameter between linear and quadratic discriminant analysis lies in the covariance matrices.[200] If both covariance matrices can be considered equal linear discriminant analysis is the best choice whereas if the matrices are different quadratic discriminant analysis is to be employed (Figure 3.14). Overall it can be said that in case of small data sets and small differences between covariance matrices LDA often is the better choice of classification method as fewer estimates need to be evaluated.

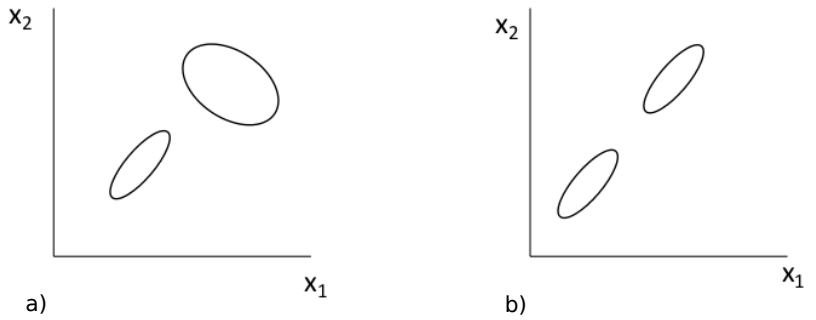


Figure 3.14: Modelled after Baldovin.[200] a) depicts analysis via QDA, b) depicts analysis using LDA

If the number of variables is comparable to the number of compounds to be estimated class covariance matrices predictions often become very unstable and the *eigenvalues* are biased on either side.[201] If the number of descriptors is higher than the number of examples of every class but lower than the total number of compounds in the data set then QDA cannot be applied because the class covariance matrix is singular. If the number of

descriptors exceeds the total number of structures in the data set neither method can be applied as both covariance matrices are singular.

Based on a study of NIR data sets Wu and colleagues compared the performance of linear discriminant analysis, quadratic discriminant analysis and regularised discriminant analysis which emphasised formerly known properties of these classification methods. QDA is not suitable unless the class sample size is significantly larger than the number of variables and is not preferable to LDA except when the covariance matrices differ.

The difference between QDA and LDA lies in the models built. QDA for example builds a model for each class. The centre of these models lies in the mean class value and its covariance matrix which means that each class is individually placed in space and volume. For LDA on the other side one model is built for the two classes together and identical population of the classes is considered. Srivastava *et al.* proposed the use of a Bayesian Quadratic Discriminant Analysis and could provide satisfactory results on ten benchmark data sets from the UCI Machine Learning Repository.[202]

Software used: lda and qda from package MASS

3.6.3 *k*-nearest neighbour (*k*NN)

The *k*-nearest neighbour approach is very easy to use and widely employed in all kinds of fields and studies.[203–206] Basically, this approach puts to use the similarity principle as it considers the specified number of neighbours of a test compound and according to the class of the selected neighbours chooses the probable class of the test compound. The starting point remains the definition of the number of neighbours taken into account and further the calculation of distances. One widely used distance measure for this purpose are euclidean distances that are calculated between the compound in question and the rest of the molecules in the training set. After the computation of a distance matrix compounds are sorted according to smallest distance between test compound and training set compounds and only the specified number of nearest neighbours is considered. The prediction of the compound in question is dependent on the average class of the nearest neighbours (Figure 3.15).

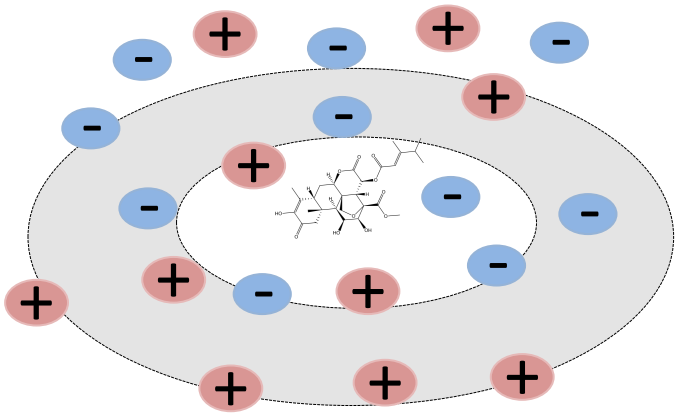


Figure 3.15: k-nearest neighbour(kNN); depicting the number of nearest neighbours taken into account.

The number of neighbours taken into account for class prediction is user defined and has to be selected individually for every data set. Sometimes the number of neighbours is fixed from the outset and in other cases the optimal number can be selected by cross-validated performance of the training set. This is comparable to parameter tuning for support vector machines.[207]

An important aspect of the relatively simple method k nearest neighbour remains the not so trivial determination of the nearest neighbours. Recent studies show that an exhaustive search for the nearest neighbours especially in a very large dataset may lead to high computational cost as the number of distances to calculate and rankings to perform grows exponentially as the number of compounds in the training set increases. Arefin and colleagues have therefore proposed parallel computing methods for large kNN problems.[208]

Another interesting problem for kNN may be due to the curse of dimensionality. This term encompasses any problem that may occur if high dimensions are reached. In this case it means that the higher the dimen-

sionality of the problem the more the distances between nearest to farthest neighbour approach each other. Consequently no ranking can be done because any difference in distances disappears. Beyer *et al.*[209] postulated that in the first 20 dimensions the contrast between distances decreases fastest. Also it has to be borne in mind that if the distance to the nearest neighbour differs only slightly from the average distance between compounds no nearest neighbour can be identified and this approach might not be useful.

3.7 The SIBAR approach

The concept of similarity based structure-activity relationship (SIBAR) has been developed in our group in 2002.[117] The primal idea that structures with similar properties also share similar activity profiles is as old as the idea of quantitative structure activity relationships itself. As the promiscuity of the ABCB1 transporter made in silico predictions highly challenging an alternative approach to common methods has been sought. A promising way to achieve this has been found in predicting ADME properties with descriptors based on similarity values. In comparison to precise structural information calculated with the usual descriptors a more general similarity based approach is used in this particular method. This approach therefore seems more suitable for targets with a high variety in their substrates and inhibitors.

In order to obtain the final SIBAR descriptors several steps have to be performed. First in line stands the selection of a suitable set of reference compounds the similarity values can be based on. In order to obtain descriptive relationship values differing similarity values to the varying members of the reference sets are of high importance. Therefore a crucial property of this reference set is high diversity as especially possible substrates of ABCB1 are characterised by high diversity and complexity. After the members of the reference sets are chosen the descriptors, for example VSA descriptors, characterising the compounds underneath have to be defined and calculated. Descriptor values are calculated for the members of the training, the test and

the reference set. After scaling of these values the final steps for SIBAR consists of the computation of similarity values for each member of the training set to each molecule of the reference set. The number of similarity values is identical to the number of members in the reference set. Similarity is calculated based on euclidean distances as this measure seemed to be the most simple and suitable approach for this purpose. Figure 3.16 illustrates the SIBAR procedure.

Based on reference compound i and a member j of the compounds in question with k molecular descriptors SIBAR values D are calculated as follows:

$$D(i,j) = \sqrt{\sum_k (X_{ik} - X_{jk})^2} \tag{3.33}$$

The received values represent the SIBAR-descriptors and they are further applied as input variables for analysis, i.e. support vector machine, random forest, etc.

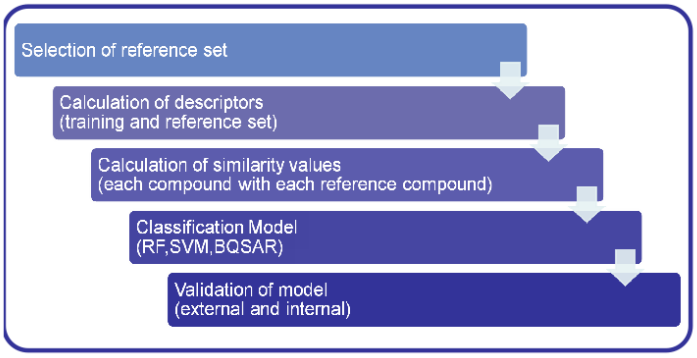


Figure 3.16: Sibar workflow

In-house results[117] employing SIBAR descriptors clearly indicated the merit of this method using a dataset of 131 compounds propafenones. With a varying set of reference compounds ranging from 10 to 40 compounds higher predictive power could be obtained than could be achieved with the molecular descriptors in pure form. This study established the usefulness of a similarity-based approach in particular with highly promiscuous targets. The results

could be undermined with another dataset consisting of intestinal absorption values for 20 diverse compounds. However, the usage of SIBAR on the benchmark set of steroids did not prove as fruitful and in this special case no statistically significant models were obtained. This was taken as further indicator that similarity values obtained via SIBAR seem to be especially suited for targets with a wide variety of ligands.

In another study of our group[210] the merit of SIBAR concerning a safe exchange of chemical information has been outlined. The business of drug development is highly competitive and pharmaceutical companies are understandably highly possessive of their possible lead structures, their targets and datasets. Nevertheless in cooperations and also to increase common knowledge the exchange of information is necessary. A possibility to achieve both aims is the employment of SIBAR as obstructive tool to disguise explicit structural information but still keep the important aspects for in silico prediction. In the mentioned study similarity searches were done based both on euclidian distances obtained from three different sets of descriptors and the SIBAR approach which in addition introduces a reference set. Though even the SIBAR approach could not create the necessary obstructive force to substantially obscure highly analogous compounds it could be shown that the SIBAR approach in comparison to simple euclidian distances significantly reduced the number of structurally analogous compounds found. It therefore represents an additional tool for exchanging chemical information without divulging an overly amount of structural information.

A further application of the SIBAR approach has been recently published by Khac-Minh Thai and Gerhard Ecker[211] on a set of hERG inhibitors. With the help of feature selection, binary QSAR and counter-propagation neural networks a number of models have been built and the applicability of SIBAR-derived models proven. The best model was derived with 11 hERG relevant descriptors and an external accuracy of 85%.

3.7.1 Global reference set

In 2000 Tudor Oprea proposed a highly interesting scheme.[212] In analogy to geography he introduced the idea of chemography where satellite structures should provide a map of the chemical space and guide the medicinal chemists. The ChemGPS was born. Instead of meridians he inserted principal properties created by principal component analysis and instead of geographical places he named molecular structures. A primal element of this plan consisted of the idea of satellite structures which are intentionally placed outside the druglike chemical space. This means that at least one property value is placed outside the common druglike values.

Two classes of compounds were selected. The first class consisted of satellite structures with extreme properties (for example glycerin, benzene, erythromycin, etc.). The second class of compounds were so-called „core structures“ and were important in order to balance the principal component model and keep the focus on the druglike chemical space. These “core structures” were taken from a list of known registered drugs with regard on their intestinal permeability properties. A number of molecular descriptors were chosen describing size, polarizability, flexibility and/or rigidity, hydrogen bonding capacity, charge and lipophilicity.

The objective of chemography is to derive a mapping device, ChemGPS, that is globally applicable and avoids extrapolation if new druglike or leadlike molecules are added to it. The initial selection of satellite structures led to a number of outliers and therefore the selection of satellite structures had to be adapted. This was done randomly and those compounds with very low probability to belong to the ChemGPS model were chosen as new satellite compounds. The aim of this project was a representation as comprehensive as possible of the druglike space so that instead of local models compounds of different origin could be compared to each other based on a globally applicable system (Figure 3.17).

As the idea of Similarity Based Structure Activity Relationship (SIBAR)

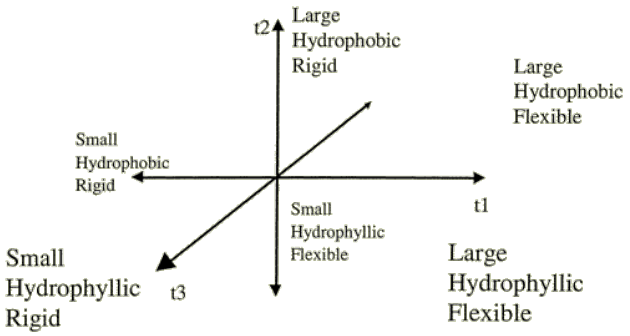


Figure 3.17: Depiction of ChemGPS principal components. Taken from Oprea.[212]

makes it necessary to generate a set of reference compounds this work provided a good starting point. In order to pursue this idea we decided to put to use the chemical database of 660 961 compounds that accompanies MOE and developed three different approaches for the selection of a globally applicable reference set.[213] The total number of molecules in the different reference sets was put to 50.

1. The 50 most diverse compounds were selected out of the MOE database on basis of the respective descriptor set (refA).

2. A pool of possible „satellite“ structures was assembled by sorting the database according to the most extreme descriptor values. Again the 50 most diverse structures were chosen out of this fund at the end of the chemical space (refB).

3. In this case the 0,1% of compounds with most extreme descriptor values were put together and again 50 molecules chosen based on maximum diversity (refC).

This procedure was repeated for every set of descriptors used to avoid any bias resulting from reference sets in the study. In the first study of this work this means that for every set of descriptors (VSA, 2D/ADME, 3D Autocorrelation) three different reference sets have been produced (Figure 3.18).

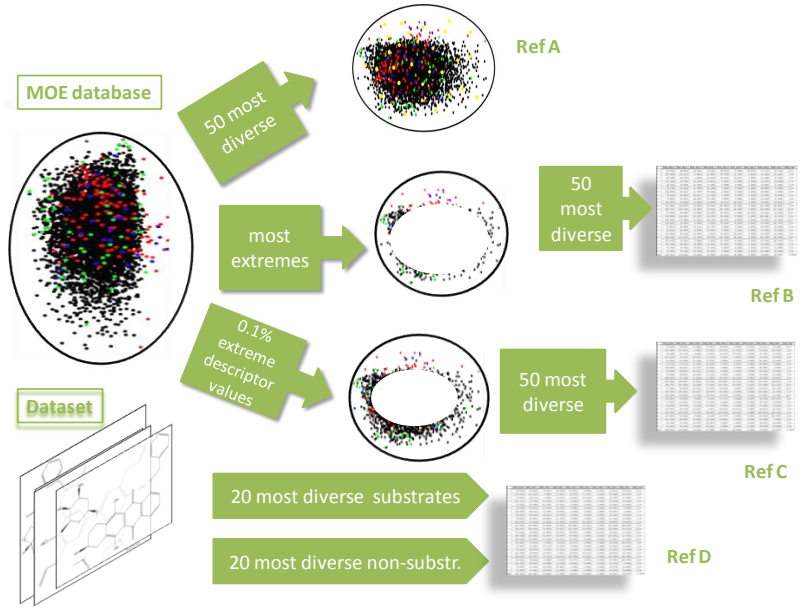


Figure 3.18: Selection of reference compounds.

3.7.2 Tailored reference set

In 2007 Barbara Zdrazil, Gerhard Ecker and colleagues conducted a study on the use of the SIBAR tool on 412 ABCB1 inhibitors with special focus on the reference set. Compared to three other more general reference sets the reference set especially tailored to the problem at hand performed best. A tailored reference set is characterised by incorporating representative compounds from the original dataset and in this case consisted of 20 molecules. Using 2D and 3D molecular descriptors partial-least-squares (PLS) analysis was performed and internally and externally validated. Results show that high diversity coupled with a relation to the target in question is highly favourable for model performance.

For this reason a fourth reference set has been compiled out of the data set used. Based on MACCS-fingerprints as implemented in MOE the 20 most diverse substrates and the 20 most diverse non-substrates have been consolidated to form a tailored reference set (Figure 3.19).

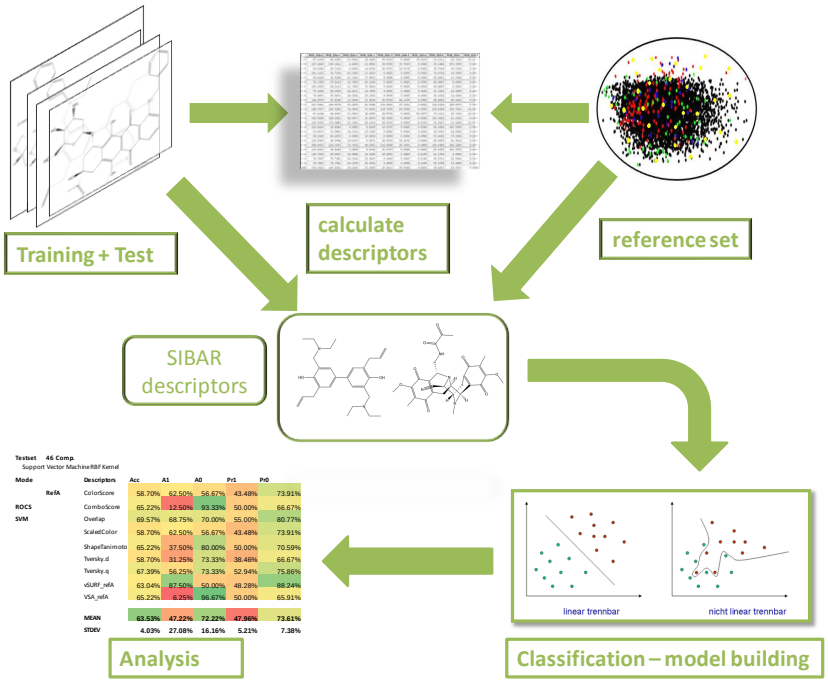


Figure 3.19: SIBAR

3.8 Descriptors

Quantitative Structure Activity Relationship (QSAR), true to its name, is dependent on the relationship between target protein and ligand. Ligand based methods rely on the detection of features all ligands for one specific target protein have in common. Therefore it is of the utmost importance to describe the molecules in a way that enables the building of a function correlating activity with specific properties or features of active molecules. These features encompass simple descriptors like the number of hydrogen bond donors/acceptors, rotatable bonds, electric charges, van der Waals Surface Area and also more complex descriptors like Volsurf descriptors or Autocorrelation descriptors. As ligand based methods or virtual screening methods deal with a high amount of compounds in one database another thing of importance is the easy and quick calculation of these descriptors. There are a number of suitable computational descriptors available and the, in our eyes, best fitting examples for this purpose have been used in this study.

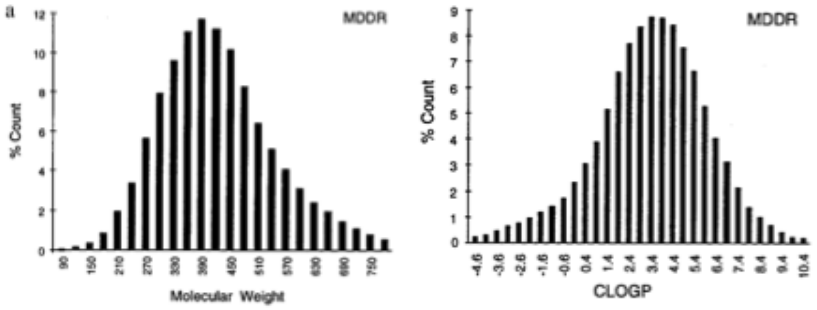
3.8.1 ADME/2D descriptors

The work of Lipinski[53] and colleagues at latest has emphasised the enormous advantages of simple rules for absorption, distribution, metabolism and elimination. In an analysis of the World Drug Index (WDI) of 50 000 drugs he selected those compounds that suggested clinical usage excluding peptides, quaternary, poly-specific drugs and compounds containing phosphatic functional groups. The resulting sum of nearly 2300 compounds was subjected to a statistical analysis. As parameters he selected the molecular weight as with increasing molecular weight the permeation decreases. Another parameter was the lipophilicity computed as logP(o/w) which as it increases is negatively correlated to permeability. Further he chose the number of hydrogen bond donors and hydrogen bond acceptors. For these two factors also negative correlation with permeability has been observed. As a consequence of this analysis he postulated the famous rule of five as coarse filter for permeability (absorption) properties of hit or lead structures in high throughput screening. Poor absorption with obvious exceptions of substrates for biological transporters, is observed if

- more than 5 hydrogen bond donors are found (sum of OHs and NHs)
- molecular weight is over 500
- logP is over 5
- more than 10 hydrogen bond acceptors (sum of Ns and Os) are present

Following Lipinski's idea of simple rules Tudor Oprea[214] published a study on general properties of druglike and non-druglike compounds. Taking various databases containing drug-like molecules like the MDL-II Drug Data Report (MDDR) and non-druglike molecules like the Available Chemical Directory (ACD) he subjected them to an analysis similar to Lipinski. The parameters he founded his analysis on were the molecular weight, ClogP, number of H-bond donors and acceptors and additionally to Lipinski the number of rings, the number of non-terminal rotatable bonds and the number of rigid bonds. As expected the ACD drugs did not exhibit the same amount of molecular complexity as the molecules of the MDDR and for this

particular purpose molecular weight and lipophilicity were too evenly distributed to allow any deduction of rules. But from the number of hydrogen bond donors, number of rings and the number of rotatable and rigid bonds some conclusion could be drawn. Druglike structures (MDDR) showed a higher probability of having rings ≥ 3, rigid bonds ≥ 18, non-terminal rotatable bonds ≥ 6 whereas non-druglike structures (ACD) often presented rings ≤ 2, rigid bonds ≤ 17, non-terminal rotatable bonds ≤ 5 (Figure 3.20).



(a) (b)

Figure 3.20: Taken from Oprea.[214] Copyright Journal of cimputer-aided molecular design. Histogram plots of distribution of molecular weight and clogP in the MDDR.

As for similarity based approaches the application of simple 2D descriptors seemed feasible a set of 88 2D descriptors was calculated using Molecular Operating Environment (MOE).

The idea was to find a set of simple descriptors able to describe the molecules adequately for input in SIBAR descriptor calculation. We wanted to explore their usability in comparison to more complex descriptors like 3D Autocorrelation descriptors or VSA descriptors.

Taking into account recent publications the following descriptor types among others have been regarded as promising and been calculated using MOE:

- The Gasteiger-Marsili Partial Charges[215] were computed as implemented

in MOE. In 1980 Gasteiger developed an iterative partial equalisation of orbital electronegativity (PEOE) which can rapidly be calculated compared with former quantum mechanical methods. The principal idea was to judge the atoms by their orbital electro-negativities. Only the connectivities of the atoms are included and thus only the topology of the molecule is regarded. As all neighbours directly connected to the atom have to be borne in mind the calculation starts with one atom and via an iterative process cycles through all of them. The rule of electronegativity, stating that the total sum of charges amounts to the overall charge of the molecule in question, is adhered to.

- The BalabanJ index has been presented in 1982 by Alexandru Balaban[216] and represents a topological index. This index is based on the average distance sum connectivity and according to the author is especially good in discriminating the „topological shape“ of chemical structures and emphasises the branching of molecules (Figure 3.21).

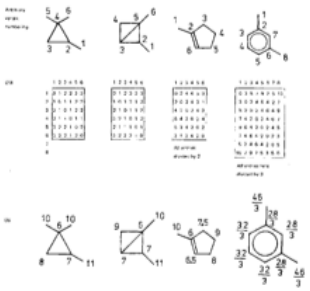


Figure 3.21: Taken from Balaban[216] depicts calculation of topological indices according to Balaban.

- Physicochemical parameters are calculated on basis of atomic contributions Labute based the van der Waals Surface Area (VSA)[217] variables on descriptors regarding atomic contributions. For that he used the descriptors presented by Wildman and Crippen[218] in 1999. They work on two kinds of proposal: the fragment based approach based on the contribution of functional groups and the atom based approach based on the contribution of each atom. Hereby the individual contributions

of logP of atoms in a molecule are summed up and thereby the total amount of logP received. They also expanded their approach to molar refractivity which commonly describes the size and polarizability of a molecule. The key to this summation is the classification of atoms to specific atom groups defined by their neighbouring groups and atoms and thereby an atom classification table was derived. The final table contained 68 basic atom types and was designed in such a way that each atom would fit only one atom type thereby avoiding confusion. The basis of this calculative effort is the 2D structure of a molecule from which the logP and the molar refractivity can be obtained. Further practical analyses have justified their approach.

- Pearlman and Smith presented the BCUT descriptors in 1998.[219] These descriptors were designed especially for diversity analysis of data sets and diversity selection. Matrices representing the hydrogen-suppressed connection table of the molecule are built. The lowest and highest *eigenvalue* are the least correlated and encompass most different information. On the diagonals four parameters (atomic charges, polarizabilities, hydrogen bond donor and acceptor abilities) are put though nominal bond-type information is kept off-diagonal of the matrix. Regarding the many different properties illuminated on the matrix scaling has to be done. The authors' intention was to position compounds in a structure-based chemistry-space and they stressed the possible unsuitability of their descripors for QSAR development. Nevertheless, these descriptors represent all features of molecular structure and for this reason they have been selected for calculation in this study of SIBAR descriptors for ABCB1 substrates and non-substrates.

- The wiener index has been introduced by Harry Wiener in[220] 1947 and represents the sum of the distances between any two carbon atoms defined by the bond between them. In order to calculate this index the number of carbon atoms on one side of any bond have to be multiplied by those on the other side. Consequently the smaller the index the more compact is the molecule. It is the oldest known topological index.

- Kier and Hall[221] Connectivity Indices are based on weighted counts of substructure fragments. They compare the molecular graph with minimal and maximal molecular graphs to detect different aspects of molecular shape.[222] The structure is taken without attached hydrogens and the sigma or valence electrons are depicted as descriptor. Further on substructures are detected and then for each fragment a connectivity index is calculated. These indices are computed from the number of heavy atom neighbours. Another type of indices are the kappa shape indices where properties of molecular shapes are encoded into three indices (Kappa values). These are also derived by counting the fragments respective their distance. Further indices are the topological and the electro-topological indices. The topological state indices are based on the topological environment of each atom with regard of all the other atoms whereas electro-topological state indices encode information about the topological and also the electronic interactions of one atom with regard on the other atoms.

A comprehensive overview of the descriptors calculated for this approach can be found in Table 3.1.

3.8.2 VSA Descriptors

Paul Labute[217] in 2000 presented a set of widely applicable descriptors. The number of descriptors alone in this study has been huge and many more are out there. It seems that for the specificity of every target individual variables are needed to describe the ligand's features satisfactorily. With his approach Paul Labute wanted to generate descriptors containing so much information that they can be used for almost every classification or regression study.

He based his ideas on the surface area of a molecule and the presumption that the shape of an atom is spherical and is defined by the van der Waals radius. This left him with the van der Waals surface area (VSA) for each atom (V_i) which summed up gave the amount of molecular van der Waals surface.

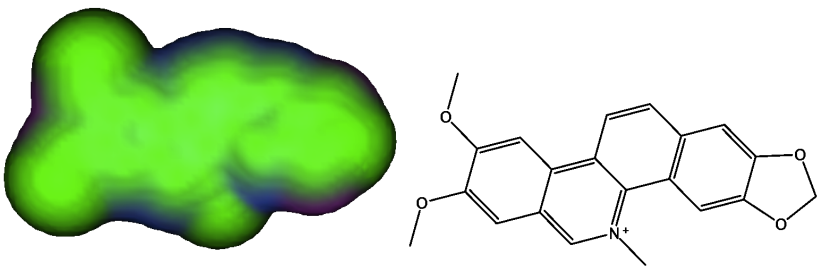
a_acc	BCUT_SMR_1	opr_nrot
a_acid	BCUT_SMR_3	PEOE_VSA_FHYD
a_aro	BCUT_SMR2	PEOE_VSA_FNEG
a_base	bpol	PEOE_VSA_FPNEG
a_don	chi0	PEOE_VSA_FPOL
a_heavy	chi0_C	PEOE_VSA_FPOS
a_nBr	chi0v	PEOE_VSA_FPPOS
a_nC	chi0v_C	PEOE_VSA_HYD
a_nCl	chi1	PEOE_VSA_NEG
a_nF	chi1_C	PEOE_VSA_PNEG
a_nH	chi1v	PEOE_VSA_POL
a_nI	chi1v_C	PEOE_VSA_POS
a_nN	chiral	PEOE_VSA_PPOS
a_nO	chiral_u	rings
a_nS	Fcharge	SlogP
apol	glob	SMR
ASA	Kier1	TPSA
b_1rotN	Kier2	vdw_area
b_ar	Kier3	vdw_vol
b_count	KierA1	VSA
b_double	KierA2	vsa_acc
b_rotN	KierA3	vsa_acid
b_single	KierFlex	vsa_base
b_triple	lip_acc	vsa_don
balabanJ	lip_don	vsa_hyd
BCUT_SLOGP_0	logP(o/w)	vsa_other
BCUT_SLOGP_1	logS	vsa_pol
BCUT_SLOGP_2	mr	Weight
BCUT_SLOGP_3	opr_brigid	weinerPath
BCUT_SMR_0	opr_nring	weinerPol

Table 3.1: Overview of calculated ADME/2D descriptors as published in Schwaha 2009[213]

$$V_i = 4\pi r_i^2 - \pi r_i \sum_{j \in B_i} \frac{r_j^2 - (r_i - d_y)^2}{d_y} \tag{3.34}$$

$$d_{i,j} = min\Big\{max\Big\{ \mid r_i - r_j \mid, b_y \Big\}, r_i + r_j \Big\} \tag{3.35}$$

Where b_y is the ideal bond length between the atoms i and j, r is the radius and B_i describes the set of all atoms bonded to atom i. d describes the distance between two centers. In this fashion the van der Waals surface area can be calculated easily by using connection table information alone, like resorting to a dictionary. As stated above the final approximate van der Waals surface area of an entire molecule is just the sum of areas calculated for each of its atoms. This assumption has been validated by calculating the van der Waals surface area with a dot-based method where each atom was surrounded by points and all the points inside any other atom were deleted. These points were then used in order to estimate the exposed surface area with the result that the approximate van der Waals surface area differed with less than 2% of individual conformations. It was decided that the approximate van der Waals surface area calculation provided reasonable results and could safely be used with the advantage of very fast calculation without any 3D information necessary (just connection table information). This means that 3D information is contained in 2D derived descriptors (Figure 3.22).



(a) Van der Waal surface (b)

Figure 3.22: Van der Waal Surface of ABCB1 substrate NSC 146397, generated with MOE.

The fundamental approach of Labute was to appoint for each atom i

in a molecule a numerical property P_p and to create a descriptor from this property value in a specific range (u,v). This descriptor is defined as the sum of the VSA contributions of each atom i with P_i in (u,v).

$$P\_VSA(u,v) = \sum_i V_i \delta(P_i \in [u,v]) \tag{3.36}$$

V_i is the atomic contribution of atom i to the VSA of the molecule. Further on n descriptors were defined by the property P.

$$P\_VSA_k = \sum_i V_i \delta_i(P_i \in [a_{k-1}, a_k]) \tag{3.37}$$

$k = 1, 2, \ldots, n$, and $a_0 < a_k < a_n$ represent the interval boundaries where the property P of the molecule is bound by $[a_0, a_n]$. In principle each VSA descriptor is defined by the sum of surface area with property P that lies in a specific range between a_0 and a_n. If the interval range contains all values then the sum of the descriptors will result in the van der Waal surface area of the molecule which means that each VSA-type descriptor represents a subunit of the van der Waal surface area in total. The properties P assigned to these descriptors should describe the physicochemical features of the molecule and were chosen to be the molar refractivity, the lipophilicity (logP) as calculated by Wildman and Crippen and Gasteiger (PEOE) method of partial charges. These methods were chosen because every one of them calculates the respective property regarding atomic contribution.

They have been implemented in Molecular Operating Environment and the interval boundaries for the VSA descriptors have been selected ensuring that the resulting intervals are evenly distributed over the database. As a result 10 descriptors have been designed characterising the logP (lipophilicity), 8 descriptors for molar refractivity and finally 14 descriptors for partial charges (PEOE).

SlogP_VSA$_k$ descriptors (10) depict hydrophobic and hydrophilic effects
SMR_VSA$_k$ descriptors (8) describe polarizability
PEOE_VSA$_k$ descriptors (14) characterise direct electrostatic interaction

That leaves us with a total number of 32 widely applicable descriptors.

In order to prove the suitability to describe molecular properties with surface area descriptors he tested the self-correlation of his descriptors. The gratifying outcome was that each of the three descriptor sets was for the most part weakly correlated with each other which is also true for all the VSA descriptors together.

In order to explore the extent the newly derived descriptor set encompasses information of other popular descriptors the 32 VSA descriptors were calculated followed by a set of 64 popular descriptors as implemented in MOE. Further on a principal components regression was calculated for each of the 64 popular descriptors presenting them as a function of the VSA descriptors. Out of these 32 showed an r^2 of 0.90 or better, 49 gave an r^2 of 0.80 or better and the last 61 showed an r^2 of 0.5 or better. These results emphasise that the developed 32 VSA descriptors contain much of the information stored in most of the 64 popular descriptors. Also, Labute tested his descriptors on seven very different targets together with the binary QSAR method with very promising results. His set of only 32 descriptors yielded good results and though maybe not able to describe one specific property in very great detail they seem perfectly appropriate to describe highly diverse compounds adequately for classification purposes.

Orthogonality is often neglected as many descriptors seem to be highly correlated with each other thereby complicating classification algorithms. It follows that good descriptors should be as orthogonal as possible. Another important topic is the relevance of the descriptors. Sometimes the described features of the molecule may be important in drug transport per se but not so very much important in the interaction between molecule and studied target. Labute refers to the fact that atomic contribution to partial charge, molar refractivity and logP are relevant in describing ligand-receptor interactions and thereby it seems that they encode information not only relevant to the receptor but also to the overall drug transport.

The here characterised descriptors have been used in a number of studies.

Zhang and colleagues[223] applied them in a combinatorial approach of several labs to a set of 159 compounds in order to develop a quantitative structure-activity relationship model for Blood Brain Barrier (BBB) permeability. The training set consisted of 144 compounds and the test set contained 15 molecules with two additional external validation sets added later on. In total six combinations of three descriptor types and two types of optimisation methods were used. The descriptors contained the MolconnZ 4.05 descriptors, MOE descriptors (including VSA descriptors) and Dragon descriptors and as methods k-nearest neighbour and linear regression support vector machine were used with feature selection. The effort resulted in six models with varying prediction accuracy and the robustness of the model was tested via Y-randomisation. The best model was achieved by k nearest neighbour and MolconnZ descriptors. An applicability domain was defined and for the external prediction sets the number of compounds satisfying the applicability domain was lower than their number in total. Nevertheless prediction accuracies scored high when the applicability domain was regarded. A consensus model was built with an external accuracy on two evaluation sets of 86.5% and 80.9% within the AD whereas the prediction accuracy dropped to 67.7% and 55.1% when outside the AD. Interpretation of the model revealed that the ten most frequently used MOE descriptors were connected to van der Waal surface area namely PEOE_VSA and SlogP_VSA which once more emphasises the suitability of van der Waal surface area based descriptors for various QSAR problems.

Thai and Ecker[145] developed a binary QSAR model of a set of 240 hERG potassium channel blockers among others using the 32 VSA descriptors with good results. Another study[224] by them compared the 32 VSA descriptors to a set of 2D descriptors derived by feature selection on a set of 285 compounds collected from the literature and proceeded with a counter-propagation neural network. The results gave a slight edge to the feature selected descriptors with an external prediction accuracy of 85% over 84% for VSA descriptors.

3.8.3 Spatial autocorrelation descriptors

These descriptors are based on the mathematical autocorrelation function

$$AC(t) = \int_a^b f(x) \cdot f(x+t)dx \tag{3.38}$$

with $f(x)$ representing a general time function and t the deferment of x. a and b define the studied interval in total.

Generally speaking autocorrelation functions describe the distribution of numerical values over an interval. This function is often used in time dependent signal processing. But it can also be used in order to describe the properties of a molecule. Gilles Moreau and Pierre Broto[225] mentioned this idea for the first time in 1980. With the atoms representing the discrete points x without taking into account any hydrogens the atom properties are described by the functions of x. Every atomic property is a standalone function. The deferment t in this case is described by the topological distances between them. Only the minimal number of bonds are counted between the atoms. This works by putting counts on every atom as discrete points in the molecule excluding the hydrogens.

The integral in this case is changed to the sum of all the function products as there is a finite number of atoms in the molecules.[226] The derived function is now

$$AC(d) = \sum_{(i,j)\in M(d)} p(i) \cdot p(j) \tag{3.39}$$

with *M(d)* defining the number of atom bonds d between atoms i and j and the property *(p)* of the atoms.[227] *AC(d)* is the Autocorrelation coefficient depending on the number of atom separating bonds. These properties can be calculated for every atom in every bond distance with the only requirement being the 2D structure of the molecule. The atomic properties on which these descriptors are based can be electronegativity, van der Waals volume, connectivity, π-functionality, hydrogen bond donor and acceptor possibilities.

In 1993 Zakarya[228] and colleagues introduced the notion of multifunc-

tional autocorrelation methods (MAM) with good results. Herein especially the atomic property of van der Waals volumes was inspected. Though generally acceptable results were obtained with this method results have been mixed in comparison with other descriptors such as the Wiener indices.

Strictly speaking every atom appears twice in the function, once as atom i and once as atom j which is regarded as negligible and can be solved by building the average:

$$\overline{AC}(d) = \frac{1}{\Delta} \sum_{(i,j) \in M(d)} p(i) \cdot p(j) \tag{3.40}$$

Δ depicts the number of products and $\overline{AC}(d)$ the average of these products not their sum. A possible problem could be that these atom properties can take positive or negative values and may cancel each other out in building the sums. As a solution the absolute values could be taken but this may put into jeopardy the information value of the descriptor.

An example of calculating the Autocorrelation descriptors is presented here (Figure 3.23).

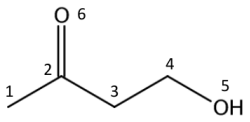


Figure 3.23: 4-Hydroxy-2-Butanon.

As atom property p we take the atom mass and consider as distance a bond length of one. Starting from atom indexed one atom masses are multiplied along the specified bond length. For illustrational purposes multiplication of each atom pair is shown only once not twice as should be done.[226]

$$AC(1) = p(1) \cdot p(2) + p(2) \cdot p(3) + p(3) \cdot p(4) + p(4) \cdot p(5) + p(2) \cdot p(6)$$
$$AC(1) = 12 \cdot 12 + 12 \cdot 12 + 12 \cdot 12 + 12 \cdot 16 + 12 \cdot 16 = 816$$
$$\overline{AC}(1) = AC(1)/5 = 163,2$$

There exist different variations of autocorrelation descriptors as for example the Moran coefficient which is defined as

$$1(d) = \frac{\frac{1}{\Delta} \sum\limits_{i,j \in M(d)} (w_i - \overline{w}) \cdot (w_j - \overline{w})}{\frac{1}{A} \sum\limits_{i=1}^{A} (w_i - \overline{w})^2} \tag{3.41}$$

with A as number of atoms in a molecule. This coefficient covers a range from -1 to $+1$. When its value is 0 this means an even distribution of properties over the molecule whereas a value of $+1$ suggest high autocorrelation. Negative autocorrelation values suggest few bearers of this property are present in distance d.

The descriptors presented above have been developed as two dimensional descriptors but auto-correlation descriptors may also be used as 3D descriptors with certain modifications.[229]

In the approach used by Wagener and coworkers which is also implemented in ADRIANA.Code[230] points are randomly scattered on the molecular surface using a preset point density. In this case the distance d no longer depicts the number of bonds between two atoms but the actual distance. Problems arise as atoms seldom lie in the exact same distance and for this reason distance intervals are determined and the coefficients calculated accordingly depending on the respective distance interval used. Hereby only the properties of the molecular surface are observed as only these are involved in protein-ligand interactions. Advantages of the autocorrelation vectors are their independence of rotation or translation as only the spatial distances are used, they mean a reduction of input information and they are very quick to calculate. However, one disadvantage remains: namely that the original input information cannot be reconstructed and conclusions such as derived by number of hydrogen bond acceptors and donors cannot be drawn from them.

In 2005 Moro and colleagues[231] investigated the principle of Molecular Electrostatic Potentials (MEP) in combination with autocorrelation vec-

tors. Hereby a unit positive charge is moved across the van der Waals surface of each molecule and measured at various points. This information was then conferred into autocorrelation vectors. The authors compared this approach favourably with descriptors obtained from the well established CoMFA method by partial least squares analysis on a set of 106 A3 adenosine receptor antagonists. Obvious advantages of this new method are that the time consuming alignment of molecules necessary for CoMFA can be omitted.

In 2012 Wang and colleagues[232] examined the advantages of 2D autocorrelation vectors versus 3D autocorrelation vectors together with other global descriptors using a support vector machine (SVM) on a set of 386 hepatitis C virus (HCV) NS5B polymerase non-nucleoside analogue inhibitors. In this study the 2D autocorrelation together with 16 global descriptors achieved the highest prediction accuracy when compared with the pairing of 3D autocorrelation and global descriptors.

In 2000 Pastor and colleagues[233] introduced another approach to autocorrelation descriptors with the 3D so called Grid Independent Descriptors (GRIND). This concept is based on molecular interaction fields (MIF) which identify so-called „virtual receptor sites“ (VRS). A grid is placed over the molecule and probe molecules are applied which can enter into atom-atom interactions like hydrogen-bonds (donor: O-probe, acceptor: N1-probe) and lipophilic interactions (DRY-probe). The differences in energy are calculated. If energy is released positive interactions take place whereas energy is needed if negative interactions occur. These energy differences are used to pick favourable regions of the molecule for binding and for measurement. The distance between these points of measurements are regarded and independent autocorrelation descriptors calculated. The difference between the original autocorrelation descriptors and GRIND descriptors however lies in the fact that no sum of products are built but only the product with the highest values for each property is kept. This has the enormous advantage that these descriptors can be traced back to the molecule if need be. Although the values of these autocorrelation descriptors can only be viewed in a so-called correlogram. These descriptors seem very promising as they are quick to calculated and the results can be traced back to the molecule.

3.8.4 VolSurf descriptors

Another approach similar to the GRIND descriptors are the VolSurf descriptors developed by Cruciani and colleagues.[234] As mentioned earlier for the GRIND descriptors which make use of molecular interaction fields other 3D molecular fields exist. The molecular electrostatic potential (MEP) for example measures the interaction energy between a unit positive charge and the unperturbed molecular charge distribution. The result is a color-coded, informative picture of the charge distribution in a molecule. Another example is the molecular lipophilicity potential (MLP) which encodes the lipophilic interactions inside the molecule. Here the molecular lipophilicity can be obtained by the logP value without the use of any probes. The MLP visualises the lipophilicity on the Solvent-Accessible Surface (SAS) by Color Coding. The GRID field as for instance used to calculate the later independent GRIND descriptors is employed and herein interaction energies investigated. Through use of probes over the surface of the investigated molecule an energy distribution map of attractive and repulsive forces is derived and can be visualised (Figure 3.24). The idea of Cruciani and colleagues however lay in the development of extracting simpler molecular descriptors called VolSurf. The starting point again are 3D molecular field maps from which the easier to handle numerical VolSurf descriptors are derived.

As the information of 2D images is encoded in pixels of different colors and patterns the 3D information of a molecule is found in its 3D molecular field map. These field maps are composed of a grid of boxes (voxels) which provide the relevant information between interacting probe and molecule like volume, surface and interaction level. These voxels are then accentuated at different energy levels and so varying images can be obtained. These images are later on extracted by VolSurf to calculate the respective surfaces and volumes. In the building phase the voxels are clustered with the aid of a shape function. For voxels inside an energy range values of 1 are conferred whereas all the others receive value 0. In this manner for example volumes are easy to calculate by summing the relevant voxels up and multiplying them by their volumes. Every change of energy level has an immediate effect on the shape

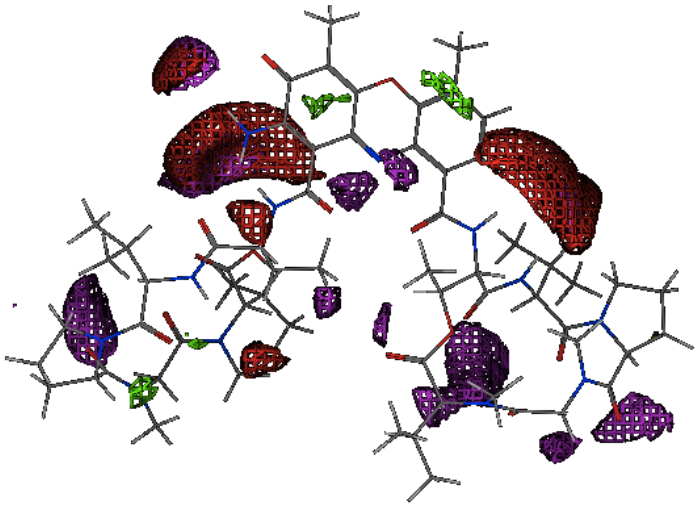


Figure 3.24: Interaction potential of ABCB1 substrate NSC3052 generated with MOE.

and size of the related volume. For VolSurf eight ranges of energy levels are employed to calculate molecular volumes and surfaces. Cruciani emphasises that the VolSurf descriptors are directly derived from the 3D molecular fields without any complicated algorithms at all. All in all VolSurf descriptors are obtained via image analysis with appropriate descriptors added depending on the 3D molecular field used.

Advantages of this approach are the easy usability and calculation speed combined with easy interpretability.

As explained above the first step consists of generating a 3D molecular field with hydrophobic and water probes and in the second step the descriptors are calculated using the information derived from the molecular fields. The descriptors encompass size and shape information, hydrophilic information, hydrophobic regions, interaction energies moments and mixed information.

In the line of similarity assessments size and shape of molecules gain more and more importance as especially for big cavities like that of P-glycoprotein

common shape could be a decisive factor. Descriptors of that ilk are the molecular volume which represents the water-excluded volume or the volume enclosed by the water-accessible surface at a repulsive value, the molecular surface (accessible surface traced out by a water probe), the ratio volume/surface which measures wrinkled surface (the smaller the ratio, the larger the wrinkled surface) and the molecular globularity which is one for perfectly spherical molecules.

Descriptors defining hydrophilic regions depict regions attracting water molecules and the capacity factors which represent the ratio of hydrophilic surface over the collective molecular surface. Descriptors of hydrophobic regions are defined by the DRY probe of the GRID field. Another set of descriptors are the INTEraction enerGY (integy) moments. They represent vectors pointing from the centre of mass to the centre of the hydrophilic regions. If this moment is high there exists a concentration of hydrophilic regions whereas at a low value hydrophilic regions are distributed quite evenly throughout the molecule or very close to the centre of mass. The last category of VolSurf descriptors are the mixed descriptors encoding local interaction energy minima derived from the interaction between water probe and molecule, energy minima distances (they describe the distances between the best three local energy minima), the hydrophilic-lipophilic balance characterising the most dominant effect in the molecule and the amphiphilic moments described by a vector pointing from the hydrophobic domain to the centre of the hydrophilic domain. The last three parameters encompassed in the VolSurf procedure are the critical packing parameters defined by the ratio between hydrophilic and lipophilic parts of the molecule, hydrogen bonding and polarizability. Polarizability is the only descriptor not computed via 3D molecular fields.

Cruciani stated that the VolSurf descriptors are hardly influenced by conformational sampling. The fact that simple 2D-to-3D conversion with energy minimisation is enough is a big advantage of VolSurf descriptors. No further molecular dynamics sampling is needed.

Crivori and colleagues[235] were the first to put to use the VolSurf descrip-

tors in a study on Blood-Brain-Barrier permeability. 110 compounds have been taken from literature and according to the afore-mentioned protocol VolSurf descriptors have been calculated. In order to investigate the importance of conformational sampling two protocols were used. In the first simple energy minimisation was performed for the dataset and in the second protocol 3D structures were fully minimised using a semi-empirical method including solvation effects. The most diverse low-energy structures were selected and VolSurf descriptors obtained. Principal component analysis (PCA) and partial least squares discriminant analysis (PLS) were performed. Similar results were obtained for both protocols and the conformational search proved to be of modest relevance and the faster method showed nearly the same results. This study emphasises the independence of VolSurf descriptors from conformational sampling.

As mentioned earlier in the first chapters (1.4.2), in 2012 Broccatelli[89] published a model for transported substrates of P-glycoprotein using VolSurf descriptors and a Naïve Bayes classifier. A dataset of 150 training set compounds and 37 external test set compounds were derived. Due to the aforementioned difficulties with definition of substrate status Broccatelli amassed a set of compounds under stringent conditions. These compounds had to be measured in the ER assay (ratio between apparent permeability from basolateral to apical direction) with MDCK-MDR1 cell lines and non-transported substances were classified as Pgp non-substrates. Models with VolSurf descriptors were done on the Orange workbench employing naïve bayes, k-nearest neighbour (kNN) and support vector machine (SVM). In Chembench random forest, support vector machine and genetic algorithm kNN (GA- kNN) were employed for six different sets of descriptors including VolSurf, Dragon with and without Hydrogens, CDK, MOE2D and MACCS. After internal 5-fold Cross-validation with at least 70% Accuracy both in specificity and sensitivity the resulting models were subjected to a feature selection method. In the end 30 descriptors were left for 6 models. In model comparison naïve bayes always outperformed the kNN models when based on the same descriptor set. The best model achieved with this approach yielded an external classification accuracy of 86% and consisted of a naïve

bayes classifier with four VolSurf descriptors. The three other good models (naïve bayes, kNN, SVM) achieved an overall external classification accuracy of 81% (naïve bayes) and approximately 78% (kNN, SVM). Feature selection searches had been done based on the naïve bayes and the kNN classifier and that could be the cause for the not so good performance of SVM. The selected descriptors may have been suboptimal for the support vector machine approach. In the Chembench approach random forest and the kNN using the genetic algorithm outperformed the SVM classifiers. Random forest coupled with VolSurf descriptors achieved an external accuracy of 84% and the genetic algorithm kNN also together with VolSurf descriptors received an external accuracy of 81%. He stated that Pgp transport seems to be enhanced by the presence of hydrogen bond donors.

Ermondi and coworkers[236] presented a study comparing VolSurf and Pentacle (formerly GRIND) descriptors on a set of non-ATP competitive Glycogen Synthase Kinase 3β (GSK-3β) inhibitors. On a set of 59 training set molecules and a test set of nine compounds both VolSurf and GRIND descriptors were calculated and feature selection applied. The final models derived via partial least squares discriminant analysis as implemented in the VolSurf or Pentacle software achieved the same external prediction accuracy of 88%. Generally GRIND based models showed better statistics but on the whole the benefit was not large enough to justify the high amount of time and care necessary when deriving GRIND descriptors. The authors suggested that for the use of binary classification studies simpler methods like VolSurf were sufficient to extract the most relevant information.

Other studies using VolSurf descriptors have been performed as well in all kinds of fields with overall good results.[237–239]

3.8.5 Rapid overlay of chemical structures

In 1993 Masek and colleagues[240] introduced a new approach of molecular shape comparisons (MSC) by optimising the overlap of volumes of two molecules. Shape comparison aims to find two or more molecules with the same spatial properties and is complicated by the conformations molecules can

adopt. In this new method suitable means are provided both for measuring the shape similarity and also to optimise the volume overlay. The issue of conformational flexibility is taken care of as low-energy conformations of the molecules are taken for pairwise shape comparison. Volume overlap comparison is made possible by regarding a molecule as sum of overlapping spherical atoms. The surface to the outside rendered by these spheres describes a molecular surface. Depending on the surface parameter chosen the overlapping volume between two molecules can be calculated. The authors of that work preferred to use van der Waals surface and the overlapping molecular volume is computed using the inclusion-exclusion principle.

This means that for two molecules (A and B) their intersection volume ($V \cap B$) can be obtained by

$$V(A \cap B) = V(A) + V(B) - V(A \cup B) \tag{3.42}$$

where $V(A \cup B)$ is the volume of A and B altogether, $V(A)$ the volume of molecule A and $V(B)$ the volume of molecule B. This principle can only be applied if the surfaces are defined by a union of spheres. Naturally the intersection volume depends on the relative position and orientation of the molecules in question. Optimal overlay of the two volumes is of the utmost importance and is accomplished by stabilising one molecule and moving the other over its surface in order to find the place of maximum volume overlay. To achieve this, complicated coordinates called Quaternions have been used as rotational variables. Concerns that the local minima would trap the volume optimisation procedure proved to be without grounds as it proceeded smoothly.

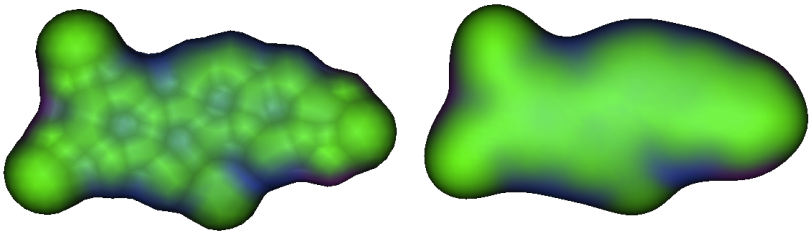
In order to find the maximum of shape overlay a search algorithm had been implemented and various volume overlays compared. Finally only a list of unique maxima of volume overlays were kept and the authors enhanced their algorithm with another feature. An option was implemented that allowed discrimination between groups with different chemical properties and thereby enabled alignment according to specific molecular properties like hy-

drogen bonding and hydrophobic interactions. Weights can be applied at will to prioritize specific features in the shape comparison process. This method though it revolutionised the idea of shape comparisons met with a few problems as the molecules were viewed as hard spheres making computing highly complicated and not always robust enough to find the global minimum.

In 1996 Grant and colleagues[241] presented the idea of Gaussian descriptors of molecular shape and moved on to prove that their idea could be favourably combined with the molecular shape comparison approach of Masek. Gaussian functions have already formerly been used to represent atomic orbitals and in this work the authors go on to describe molecular shape by regarding each atom as a Gaussian function and thereby derive its molecular volume. This work had an important impact on the birth of ROCS[100] because in this manner all needed volumes could be calculated easily and speedily due to the smoothness of Gaussian functions. Shape matching in Grant's work was done similar to Masek and coworkers[240] as the intersection volume of two molecules was calculated by rigidly translating and rotating one of the molecules with respect to the other.

In Gaussian description of molecular shape only two parameters evolve for describing each atom. One is the atomic radius which can be obtained by a list of Connolly[242] and a Gaussian weight which has been assigned a universal value of 2.7. The authors could prove that the Gaussian overlap model behaved correctly and compared with the hard spheres algorithm presented by Masek. Four possible initial starting points were deemed to be enough for finding the global minimum for each molecule. Matching methods based on hard spheres are associated with slow computation due to complicated mathematical formulae used whereas the Gaussian technique couples fast computation time with more physically realistic descriptors of a molecule and convergence to the global minimum is on the whole possible (Figure 3.25).

The final stepping stone for the rapid overlay of chemical structures (ROCS)[100] program of openeye as we know it has been provided in 2005 by Rush and colleagues.[243] On a set of inhibitors of the ZipA-FtsZ protein-



(a) Connolly surface (b) Gaussian surface

Figure 3.25: Surfaces of ABCB1 substrate NSC146397, generated with MOE.

protein interaction the authors defined overlap of shape O_{AB} between two molecules A and B as

$$O_{A,B}(\vec{q}^A, \vec{q}^B) = \int\int\int \chi^A(\vec{r}, \vec{q}^A)\chi^B(\vec{r}, \vec{q}^B)d\vec{r} \tag{3.43}$$

with r representing the position in space, q is the set of variables that provide orientation and position information and χ is the „characteristic volume“ function. Using a set of reduced Gaussians did not appear to be a problem and further on the shape distance $D_{A,B}$ could be formulated as

$$D_{A,B} = \sqrt{O_{A,A} + O_{B,B} - 2O_{A,B}} \tag{3.44}$$

which is a true metric and predicts that shape is an intrinsic property. From the received overlays one can also calculate the similarity of the overlay in terms of the Tanimoto index $T_{A,B}$ like

$$T_{A,B} = \frac{O_{A,B}}{O_{A,A} + O_{B,B} - 2O_{A,B}} \tag{3.45}$$

Shape Tanimoto is 1 if two shapes are identical and 0 if the two shapes are completely different with zero overlap. ROCS uses matches based only on volume overlap of optimally aligned molecules and the outcome should open up a new dimension of scaffold hopping. It is important to generate all sorts of conformations for each molecule so that shape overlay algorithm is not dependent on a single conformation which would restrict the results.

For Rush's analysis the whole structure of the query molecule and also a substructure were taken to screen a database for molecules similar in shape and the resulting hits were then overlaid with the query molecule and the intermolecular van der Waals energy computed. If positive interaction energy resulted these compounds were discarded. As no functional groups influenced the similarity search obvious clashes between protein target and query ligand had to be avoided. That method provided the authors with two new scaffolds and resulted further on in three possible lead structures for the protein target.

Many applications of ROCS have been published to this date highlighting the efficiency of this approach for virtual screening. In 2007 Hawkins and colleagues[244] observed the performance of ROCS in comparison with a group of docking programs to test their applicability as virtual screening tools. The published results emphasised that ROCS performed at least as well if not better as the docking programs tested and represents a real alternative for virtual screening. Other similar studies have been published.[245,246] In 2009 Haque and colleagues[247] proposed a scheme to accelerate the parallel performances of ROCS using PAPER. PAPER is an open-source implementation of Gaussian molecular shape overlay and the authors demonstrate one or two order-of-magnitude speedups if PAPER is used.

3.8.5.1 The program ROCS by Openeye

As explained above based on a query molecule ROCS performs shape overlays in search of the maximum overlap between query molecule and database molecule. Usually it is sufficient if the query molecule is presented as single structure whereas conformers of the molecules in the database have to be built in order not to artificially lessen the number of possible shape overlays. The conformations can be generated using the openeye product OMEGA.[248] At first the centres of mass of query and candidate are evaluated and alignment follows based on their principal components of inertia. Subsequently volume overlaps are optimised by application of a solid-body optimisation algorithm and the resulting hits ranked according to similarity indices like

the Tanimoto index or the ColorScore.[249]

Shape similarity S_1 as defined in the ROCS[100] manual is considered as

$$S_1 = \int \mid f(x,y,z) - g(x,y,z) \mid dV \tag{3.46}$$

where f and g are different characteristic functions. If the integral is zero then f and g are the same shape and the larger the integral the more difference between the two shapes.

The fundamental equation for shape comparison in ROCS is

$$S_{f,g} = I_f + I_g - 2O_{f,g} \tag{3.47}$$

where the I terms are the self-volume overlaps of each molecule while the O term is the overlap between those two functions.

Different similarity indices are calculated by the program enabling the user to choose his own preferred similarity measure. All of them calculated for each molecule in the database are written in the report file generated by ROCS.

The **Shape Tanimoto** represents the quantitative overlap between two structures and is calculated as shown above.

The **Tversky index** is another measure of similarity.

$$Tversky_{f,g} = \frac{O_{f,g}}{\alpha I_f + \beta I_g} \tag{3.48}$$

The Tversky index depends on which molecule's self-overlap has the α pre-factor. This pre-factor puts weights on the contribution of the molecule it is based on. The larger the α pre-factor the more important is the contribution of "its molecule". Two indices are calculated with the α pre-factor on the query molecule and another index the other way round with the α pre-factor on the database molecule. Tversky indices are used mostly to calculate the similarity between bit-vector fingerprints. This index can be larger than 1.0 since the overlap O_f, g can be larger than a molecule's self overlap I_f.

Tversky.d with the database molecule as the main self-overlap withβ= 0.95 and Tversky.qx with the query molecule as the main self-overlap term with $\alpha = 0.95$

ColorScore: There is no upper bound on this score due to two different force fields (Implicit MillsDean and ExplicitMillsDean) used that determine chemical functions and their interactions. Six different types of molecules can be characterised consisting of hydrogen-bond donors, hydrogen-bond acceptors, hydrophobic, negatively, positively and aromatic moieties and this score depicts their distribution in the two molecules. Over time the usefulness of implementing such an index has been perceived where combined with shape similarity also the features of the molecules can be taken into account. This score is coded by looping over all the color atoms in the query molecule and summing the single best color interaction with the hit molecule. Implemented in the force field also is a basic pK_a model with pH of 7 so that no protonation of molecules has to be done and charges are assigned automatically.[249] This leads to scores that quantify the match between corresponding features as normally seen in pharmacophore matching programs[100] but its maximum value represents the self-color of the query molecule.

Scaled Color: Via scaling the values of the colorscore may be seen without considering the forcefield used. Here the hit's actual score value is taken and then divided by the ColorScore of the query molecule against itself. The query self-color is at a theoretical maximum and the scaled ColorScore is restricted between 0 and 1.

ComboScore: This score is a combined value of the color-encoded molecular features together with the ShapeTanimoto index. It simply equally summarises the two other scores and accordingly its value can lie between 0 (no overlap) and 2 (complete overlap).

Subtan: can be calculated additionally but is not included in the default options. This index is defined by taking the positions of the query

and database molecules at the final overlay and removing all database atoms greater than 1.5 Å from any query atom. Shape Tanimoto calculations are subsequently performed and the resulting index is called SubTan index. As a consequence scores for small queries run against a database with larger molecules can be raised.

Overlap: the absolute overlap of volume between both molecules.

3.9 Experimental

3.9.1 Data set

The original data set consisted of 240 compounds containing 120 non-substrates and 110 substrates. The data set and the reference sets were washed using moe.wash function as implemented in Molecular Operating Environment (MOE[137] 09.2007). The washing function with default values in MOE consists of various filter procedures like the filtering of water molecules, counter ions, oxygen ions, simple acids and bases, common solvents and common salts for curation of the data set.

Structures have been minimised using the function energy.minimize as implemented in MOE 09.2007 and MOE 10.2008. According to the manual, this function tries to find a set of atomic coordinates that are coupled with a local minimum of the molecular energy function. Hereby large scale non-linear optimisation techniques are conducted to calculate a conformation where forces on atoms are as low as possible. First of all a test of convergence is done which is followed by the computation of the search direction and then the further step sizes are computed. Convergence is found when any of the three following conditions are simultaneously satisfied. They comprise a root mean square gradient test whose value was set to 0.01, an iteration limit test and a progress test. The program selects one of three methods for energy minimisation which consist of steepest descent method, conjugate gradient

method and truncated Newton. Steepest descent is only used if the gradient is very high whereas the conjugate gradient method is employed if the gradient is sufficiently small but does not easily converge. Once the energy gradient is reasonable the Truncated Newton method is used until convergence. The default values were taken with the forcefield MMFF94 (Manual of MOE[250]).

Partial charges have been calculated using the PEOE charges as suggested by Gasteiger and Marsili[215] and implemented in MOE.

After that MACCS fingerprints were calculated for the original dataset as implemented in MOE 09.2007. MDL 166 Fingerprints (MACCS) encode molecular features into binary „keybits“.[251] An ordered collection of keybits is pronounced a „keyset“. The MDL keysets used in this study contain 166 keybits (MACCS) and have been applied multiple times in molecular modelling. They are based on the 2D or 3D structures of molecules and present numerical values describing many structural features of a molecule like for example the number of atoms, bonds, etc. A keybit is defined by nine numbers. The first four of these describe various properties of the molecules while the last five numbers determine the keybits set by the specific features like heteroatoms, halogens, aromatics etc. They were originally designed for searching databases.[252] In principal if a certain functional group or fragment is present in a molecule the respective bit position is set „1“ and „0“ if this specific functional group is missing. When applied in similarity searches the results have been quite promising.[253] A possible drawback may be that fingerprints of larger and more complex molecules generate fingerprints of higher density than less sophisticated examples.

After calculation of the MACCS fingerprints the diverse.subset function implemented in MOE 09.2007 was employed to extract the 20 most diverse substrates and the 20 most diverse non-substrates for the tailored reference set.

Diverse.subset

This function defines ranks for each entry of the database based on maximum distance between the compounds and can be used after clustering or for the dataset by and large. For this purpose the distance between each ranked and unranked entry is determined and the minimum distance for each unranked entry to each ranked entry is calculated. The one compound deemed farthest away is the one with the largest minimum distance. In order to calculate the distance euclidian distance measures are used on basis of descriptors or Tanimoto similarities when calculated on basis of fingerprints as has been done in this case (Manual of MOE[250]).

3.9.2 Validation

Internal validation

In this study internal validation was used on all models. In case of naïve Bayes classification leave-one-out cross-validation as implemented in MOE 09.2007 and 10.2008 was employed. For the SVM method using WEKA 3.5.7 leave-ten-out as implemented in the software was used. For all other methods an in-house script of our group was used to generate leave-ten-out models and therefore optimally validate the models internally. As this model was intended primarily for exploration of different classification methods and especially for exploration of similarity based approaches regarding ABCB1 no additional Y-randomisation procedure was deemed necessary.

External validation

For the first study 80% of the most diverse compounds based on MACCS fingerprints and diverse.subset as implemented in MOE were exported into the training set whereas the remaining 20% comprised the external test set. The resulting training set contained 72 substrates and 80 non-substrates and the external test set consisted of 18 substrates and 30 non-substrates.

3.9.3 Classification

First study

Further on 32 VSA descriptors were calculated for training, test and each reference set and after scaling the relevant SIBAR values computed using one reference set after the other. The same procedure was applied for determination of 88 2D/ADME descriptors and the subsequent SIBAR values. In order to receive the relevant Autocorrelation descriptors the data sets and each reference set were imported into ADRIANA Code[230]by Molecular Networks and 84 3D Autocorrelation descriptors and the respective SIBAR values received. Classification studies encompassed binary QSAR as implemented in MOE 09.2007 and support vector machine as implemented in the WEKA[158] workbench version 3.5.7 with the help of the implemented Grid Search function.

Second study

The aim of the second study was to further explore shape similarity approaches combined with the SIBAR method. Therefore OMEGA[248] version 2.2.1 from OpenEye Scientific Software was employed for 3D conformer generation as input into the ROCS database. OMEGA is a very reliable software package and encourages a limit of a maximum of 20 rotatable bonds on structures for reliable conformation generation. Therefore the original dataset was reduced and that resulted in a total of 233 compounds. 40 of these compounds were taken as reference set D. 75% of the remaining compounds were contained in the training set with 69 substrates and 78 non-substrates. The test set consisted of the residual 25% with 16 substrates and 30 non-substrates.

The structures were washed using the wash function implemented in MOE 10.2008,[137] energy-minimised and PEOE charges added. Again 32 VSA descriptors were calculated for training and test set, scaled and the respective

SIBAR values derived. Previous studies[235] show simple energy minimisation produces similar results to more elaborate conformational sampling and for that reason simple energy minimisation as implemented in MOE 10.2008 was used before the calculation of 76 VolSURF descriptors and SIBAR values. ROCS[100] (rapid overlay of chemical structures) version 2.3.1 from OpenEye Scientific Software was employed with the training and test set as screening database. The reference sets were taken as query structures and the resulting parameter values taken from the report file. These parameter values were then used individually as input descriptors for further classification.

Classification methods of the second study have been random forest, support vector machine, binary QSAR, linear discriminant analysis and quadratic discriminant analysis.

Random forest classification was done using the randomforest[196] package of the R project version 2.9.1.[159] The support vector machine including the tune function was derived from the e1071[162] package of the R project version 2.9.1. Linear and quadratic discriminant analysis was performed using the MASS[254] package from the R project version 2.9.1.

The k-nearest neighbour approach was achieved using the kNNcat[255] package of the R project version 2.9.1. The nearest neighbours were set from 1 to 5 and the best model selected. Class prediction was done by majority vote of the specified nearest training set neighbours based on the ROCS ComboScore.

4

Results and Discussion

The aim of this study has been the exploration of shape similarity methods in combination with an enormously promiscuous protein like ABCB1 (P-glycoprotein). Especially the in-house developed similarity based descriptors have been the focus of this study. In combination with shape similarity and 3D based methods we wanted to test the applicability of this approach and maybe present a universally applicable modelling approach that can be of further use either in industry or academic research. For this reason the aforementioned classification methods and descriptors have been employed with following results.

The assessment of the model performance in this study unless otherwise stated is based on results using the external test set for validation purposes. The measurements of performance used in this study comprise overall Accuracy (A), Accuracy on substrates (A1 – Sensitivity), Accuracy on non-substrates (A0 – Specificity), Precision on substrates (Pr1) and Precision on non-substrates (Pr0). Another evaluating tool for performance of machine learning methods is the Matthews Correlation Coefficient (MCC) which has been used in the second study. An overview of how these values are calculated is given below

Overall accuracy:

$$A = \frac{TP + TN}{TP + FP + TN + FN} \tag{4.1}$$

Accuracy on substrates (sensitivity):

$$A1 = \frac{TP}{TP + FN} \tag{4.2}$$

Accuracy on non-substrates (specificity):

$$A0 = \frac{TN}{TN + FP} \tag{4.3}$$

Precision on substrates:

$$Pr1 = \frac{TP}{TP + FP} \tag{4.4}$$

Precision on non-substrates:

$$Pr0 = \frac{TN}{TN + FN} \tag{4.5}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4.6}$$

4.1 Methods

4.1.1 Binary QSAR

The method developed by Labute, as mentioned previously, combines easy handling with persuasive results and has been proven successful in our group with respect to the hERG[145] (human Ether-à-go-go-Related Gene) encoded channel.

4.1.1.1 Autoqsar

In a first approach in order to get a feeling for the potential model that can be built using binary QSAR the script autoqsar,[256] available from the SVL-exchange server has been used. It has been developed by the company Ryoka Systems Inc. and put online for free use. This script imports the original file of molecules into a new database and first starts the washing function of MOE, then further adds partial charges and energy minimises the molecules

in the database accordingly. Descriptors can be specified for either use or exclusion and then the preferred classification method is selected. There are three methods on disposal including principal components regression, partial least squares and binary QSAR based on Bayes theorem. Further on the user can specify whether to use a fixed number of descriptors for the final model or to perform feature selection with the smallest number of descriptors for the best model. The maximum number of principal components used in this analysis can also be fixed. In this study the number of components has not been specified and the best model with the smallest number of descriptors has been opted for which is identified via leave-one-out cross validation. Another aspect of the autoQSAR script is that it optimises the model regarding the accuracy on substrates.

Hillebrecht and Klebe[257] have used this method for a quick overview and as a feature selection method as they compared the applicability of established 3D QSAR methods like CoMFA, CoMSIA in database screening with conventional 2D QSAR models based on MACCS fingerprints or VSA descriptors. They established a protocol for 3D database screening and showed that 3D methods perform either comparable or better than MACCS fingerprints or VSA descriptors. In this way the script has also been used in a very interesting study of Zhu, Tropsha, Fourches and colleagues[258] where each of six academic groups built QSAR models based on the same dataset with varying methods. The methods ranged from kNN to SVM and neural networks. All the models were evaluated with the same external test sets and a consensus model built. The autoqsar method in this instance also was used as a feature selection procedure and the most important descriptors selected accordingly.

The best result of this study using this script has been given by the SIBAR model using 3D Autocorrelation vectors together with reference set A. The overall external accuracy amounted to 75% with 50% accuracy on substrates, 90% accuracy on non-substrates and 75% each on precision on substrates and non-substrates. Though this is not bad an accuracy on substrates of only

Descriptor	A	A1	A0	Pr1	Pr0
3DAuto refA	**75,00**	**50,00**	**90,00**	**75,00**	**75,00**
VSA refA	68,75	27,78	93,33	71,43	68,29
2D refA	47,92	33,33	56,67	31,58	58,62
3DAuto refB	64,58	50,00	73,33	52,94	70,97
VSA refB	68,75	50,00	80,00	60,00	72,73
2D refB	70,83	33,33	93,33	75,00	70,00
3DAuto refC	64,58	50,00	73,33	52,94	70,97
VSA refC	66,67	22,22	93,33	66,67	66,67
2D refC	66,67	83,33	56,67	53,57	85,00
3DAuto refD	39,58	100,00	3,33	38,30	100,00
3DVSA refD	62,50	55,56	66,67	50,00	71,43
2D refD	60,42	88,89	43,33	48,48	86,67
only 3DAuto	**64,58**	**50,00**	**73,33**	**52,94**	**70,97**
only VSA	64,58	27,78	86,67	55,56	66,67
only 2D	54,17	11,11	80,00	25,00	60,00

Table 4.1: Overview over models built using binary QSAR with the script autoqsar. Bold letters indicate the best models. A – overall accuracy, A1 – accuracy on substrates, A0 – accuracy on non-substrates, Pr1 – precision on substrates, Pr0 – precision on non-substrates. 3D Auto -3D Autocorrelation descriptors, VSA – VSA descriptors, 2D – 2D/ADME descriptors, refA – reference set A, refB – reference set B, refC – reference C, refD – reference D, only – pure descriptors used.

50% is far away from an optimal model. Nevertheless this script presents a usable approach for a quick overview over one's data. The best model using the „pure“ descriptors was given by the 3D Autocorrelation descriptors with an accuracy of nearly 65%, nearly 50% accuracy on substrates, nearly 74% accuracy on non-substrates and 53% and 71% on precision on substrates and non-substrates. This instance shows that overall accuracy considered alone without the other parameters can be deceiving as an overall accuracy of 65% may not be perfect but acceptable whereas an accuracy on substrates of 50% means more or less random prediction. The rest of the models mostly lie

around 65% overall accuracy with varying values of sensitivity, specificity and precision. Interestingly, though optimisation on substrates is the chief driving force of the script the resulting accuracies mostly lay between 25 to 50% with the occasional outlier in both sides (Table 4.1).

4.1.1.2 Binary QSAR: no restraint on principal component

The following models have been built using the normal binary QSAR procedure as implemented in MOE.[137] A predicted activity lower than 0.5 results in a non-substrate verdict opposed to higher than 0.5 predicted activity which bins the compound as substrate. In the first binary model the components have not been restrained to a specific number in order to get an insight how much the restriction of principal components would weigh with regard to prediction accuracy and modelling success. The accuracies of the resulting models have been mixed. The best result was achieved using the 2D descriptors with reference set C with an overall accuracy of nearly 73%, an accuracy on substrates of 67% and accuracy on non-substrates of 77% with a precision of 63% and 79% respectively. The best model using the pure descriptors alone gave an overall accuracy of nearly 69% with once again only 33% accuracy on substrates, 83% accuracy on non-substrates and 80% and 67% each precision.

On the whole the results have been mixed as the overall accuracy ranges from 40% to nearly 73%. Generally it was observed that VSA descriptors performed best with regard to overall accuracy, followed by the 2D descriptors and least performance was observed with 3D Autocorrelation descriptors. Nevertheless when calculating the mean over overall accuracy, sensitivity, specificity and the two precision values the 3D Autocorrelation showed the least deviation from the mean regarding sensitivity and specificity. Thus though not achieving as high an overall accuracy as the VSA descriptors they showed the more stable performance (Table 4.2).

Descriptors	A	A1	A0	Pr1	Pr0
3DAuto refA	64,58	44,44	76,67	53,33	69,70
VSA refA	64,58	16,67	93,33	60,00	65,12
2D refA	60,42	16,67	86,67	42,86	63,41
3DAuto refB	54,17	38,89	63,33	38,89	63,33
VSA refB	68,75	22,22	96,67	80,00	67,44
2D refB	58,33	22,22	80,00	40,00	63,16
3DAuto refC	60,42	55,56	63,33	47,62	70,37
VSA refC	68,75	33,33	90,00	66,67	69,23
2D refC	**72,92**	**66,67**	**76,67**	**63,16**	**79,31**
3DAuto refD	37,50	77,78	13,33	35,00	50,00
3DVSA refD	70,83	66,67	73,33	60,00	78,57
2D refD	37,50	100,00	0,00	37,50	0,00
only 3DAuto	54,17	44,44	60,00	40,00	64,29
only VSA	**68,75**	**22,22**	**96,67**	**80,00**	**67,44**
only 2D	64,58	33,33	83,33	54,55	67,57

Table 4.2: Overview over models built using binary QSAR with no limit in principal components. Bold letters indicate the best models. A – overall accuracy, A1 – accuracy on substrates, A0 – accuracy on non-substrates, Pr1 – precision on substrates, Pr0 – precision on non-substrates. 3D Auto -3D Autocorrelation descriptors, VSA – VSA descriptors, 2D – 2D/ADME descriptors, refA – reference set A, refB – reference set B, refC – reference C, refD – reference D, only – pure descriptors used.

4.1.1.3 Binary QSAR: number of principal components restrained to a maximum of 15

The benefit of principal component analysis lies in the fact that the most important elements of molecular descriptors are contained in the first few principal components and redundant or unnecessary information is eliminated or banned to the last components. Therefore a restraint on the principal components used in this method has been placed. The final models have been restricted to the use of a maximum of 15 principal components.

The VSA descriptors have been announced as orthogonal descriptors and with the use of SIBAR descriptors the number of 15 components seemed justified. Interestingly, though autoqsar supposedly searches for the best model via leave-one-out cross-validation on the fly, the models achieved with the restraint to 15 principal components and normal binary QSAR procedure were more consistent regarding sensitivity and specificity than the autoQSAR method. The best model was given by 2D SIBAR descriptors using reference set C with an overall accuracy of nearly 73%, an accuracy of substrates of 67%, an accuracy on non-substrates of 77%, 63% and 79% precision on substrates and non-substrates. Though the best overall accuracy has been achieved with 75% by the VSA SIBAR as well as the 2D SIBAR descriptors using reference set A the residual accuracies and precisions did not measure up. Especially for VSA SIBAR and reference set A accuracy on substrates was a disappointing 39%, with 97% accuracy on non-substrates and 88% and 73% precision on substrates and non-substrates.

The best model using „pure“ descriptors returned an overall accuracy of 65% with an accuracy on substrates of 33%, accuracy on non-substrates of 83%, precision on substrates 55% and precision on non-substrates 68% using the 2D descriptors (Table 4.3).

A general observation throughout all models so far presented is the fact that prediction of possible substrates remains a very challenging task for this classification method. The precision on non-substrates and the accuracy on non-substrates mostly by far outperforms the precision on substrates and their respective accuracy.[213]

The second part of the study was comprised of matching the performance of shape-based methods like ROCS with conventional 3D descriptors like VolSurf descriptors and easily calculable VSA descriptors which also contain hints of 3D information. As described earlier not all the compounds used for the earlier study could be held on to due to problems regarding con-

Descriptor	A	A1	A0	Pr1	Pr0
3DAuto refA	56,25	33,33	70,00	40,00	63,64
VSA refA	**75,00**	**38,89**	**96,67**	**87,50**	**72,50**
2D refA	**75,00**	**55,56**	**86,67**	**71,43**	**76,47**
3DAuto refB	66,67	61,11	70,00	55,00	75,00
VSA refB	62,50	22,22	86,67	50,00	65,00
2D refB	70,83	44,44	86,67	66,67	72,22
3DAuto refC	54,17	55,56	53,33	41,67	66,67
VSA refC	72,92	38,89	93,33	77,78	71,79
2D refC	**72,92**	**66,67**	**76,67**	**63,16**	**79,31**
3DAuto refD	60,42	11,11	90,00	40,00	62,79
VSA refD	62,50	16,67	90,00	50,00	64,29
2D refD	66,67	55,56	73,33	55,56	73,33
only 3DAuto	56,25	66,67	50,00	44,44	71,43
only VSA	62,50	27,78	83,33	50,00	65,79
only 2D	**64,58**	**33,33**	**83,33**	**54,55**	**67,57**

Table 4.3: Overview over models built using binary QSAR with limit to 15 principal components. Bold letters indicate the best models. A – overall accuracy, A1 – accuracy on substrates, A0 – accuracy on non-substrates, Pr1 – precision on substrates, Pr0 – precision on non-substrates. 3D Auto -3D Autocorrelation descriptors, VSA – VSA descriptors, 2D – 2D/ADME descriptors, refA – reference set A, refB – reference set B, refC – reference C, refD – reference D, only – pure descriptors used.

formational sampling. The problematic compounds were removed from the datasets and thereby the training set slightly reduced. Therefore the binary QSAR model had to be done again with the same parameters as before.

With the different descriptors apart from VSA descriptors the results were again mixed. The most reliable descriptor derived from ROCS was the ColorScore descriptor. This descriptor encodes the distribution of functional groups in the two compared molecules thereby placing importance on the features not only the shape of the molecule. According to the manual

of ROCS[259] this value is comparable to pharmacophore matching. Though the SIBAR ColorScore showed the best general performance with overall accuracy ranging from 67% to 72% for the three satellite reference sets for reference set D the ComboScore and the Overlap obtained the best specific results (Table 4.4).

The best overall result therefore was achieved using reference set D with the parameter Overlap with an overall accuracy of 80%, an accuracy of substrates of 69%, an accuracy of non-substrates of 87% with a precision of substrates of 73% and 84% for non-substrates. This was the best model so far achieved with binary QSAR methods (MCC = 0.56). The best result using the VolSurf descriptors was given using reference set C with an overall accuracy of 72% but with a disastrous accuracy on substrates of 44%, the accuracy on non-substrates remained 87%. The VSA descriptors did not perform very well in this constellation with a best performance of overall accuracy of 67% using reference set D. The best performance with „pure“ descriptors was given with the VolSurf descriptors with an overall accuracy of nearly 61% an accuracy on substrates of 37,5%, an accuracy on non-substrates of 73% and a precision of 43% and 69% on substrates and non-substrates.[260]

Table 4.4: Results of binary QSAR of study 2

Testset	46 Comp. BQSAR Descriptors	A	A1	A0	Pr1	Pr0
refA	ColorScore	71,74	50,00	83,33	61,54	75,76
	ComboScore	63,04	50,00	70,00	47,06	72,41
	Overlap	67,39	18,75	93,33	60,00	68,29
	ScaledColor	65,22	56,25	70,00	50,00	75,00
	ShapeTanimoto	54,35	12,50	76,67	22,22	62,16
	Tversky(d)	60,87	43,75	70,00	43,75	70,00
	Tversky(q)	69,57	62,50	73,33	55,56	78,57
	vSURF refA	63,04	0,00	96,67	0,00	64,44
	VSA refA	58,70	0,00	90,00	0,00	62,79

Table 4.4: Results of binary QSAR of study 2

Testset	46 Comp. BQSAR Descriptors	A	A1	A0	Pr1	Pr0
refB	ColorScore	71,74	56,25	80,00	60,00	77,42
	ComboScore	60,87	25,00	80,00	40,00	66,67
	Overlap	58,70	43,75	66,67	41,18	68,97
	ScaledColor	52,17	43,75	56,67	35,00	65,38
	ShapeTanimoto	54,35	25,00	70,00	30,77	63,64
	Tversky(d)	54,35	18,75	73,33	27,27	62,86
	Tversky(q)	67,39	43,75	80,00	53,85	72,73
	vSURF refB	73,91	31,25	96,67	83,33	72,50
	VSA refB	54,35	6,25	80,00	14,29	61,54
	10ADME refB	63,04	18,75	86,67	42,86	66,67
refC	ColorScore	67,39	62,50	70,00	52,63	77,78
	ComboScore	54,35	25,00	70,00	30,77	63,64
	Overlap	60,87	50,00	66,67	44,44	71,43
	ScaledColor	71,74	50,00	83,33	61,54	75,76
	ShapeTanimoto	63,04	37,50	76,67	46,15	69,70
	Tversky(d)	56,52	50,00	60,00	40,00	69,23
	Tversky(q)	65,22	25,00	86,67	50,00	68,42
	vSURF refC	**71,74**	**43,75**	**86,67**	**63,64**	**74,29**
	VSA refC	56,52	0,00	86,67	0,00	61,90
refD	ColorScore	60,87	37,50	73,33	42,86	68,75
	ComboScore	73,91	56,25	83,33	64,29	78,13
	Overlap	**80,43**	**68,75**	**86,67**	**73,33**	**83,87**
	ScaledColor	65,22	43,75	76,67	50,00	71,88
	ShapeTanimoto	60,87	25,00	80,00	40,00	66,67
	Tversky(d)	60,87	25,00	80,00	40,00	66,67

Table 4.4: Results of binary QSAR of study 2

Testset	46 Comp. BQSAR Descriptors	A	A1	A0	Pr1	Pr0
	Tversky(q)	63,04	31,25	80,00	45,45	68,57
	vSURF refD	67,39	12,50	96,67	66,67	67,44
	VSA refD	**67,39**	**37,50**	**83,33**	**54,55**	**71,43**
	10ADME refD	65,22	50,00	73,33	50,00	73,33
pure	**only vSURF**	**60,87**	**37,50**	**73,33**	**42,86**	**68,75**
	only VSA	60,87	12,50	86,67	33,33	65,00
	only 10ADME	58,70	37,50	70,00	40,00	67,74

Table 4.4: External results of binary QSAR restricted to 15 components. Best results are highlighted in bold letters. A – overall accuracy, A1 – accuracy on substrates, A0 – accuracy on non-substrates, Pr1 – precision on substrates, Pr0 – precision on non-substrates, vSURF – VolSurf descriptors,, refA – reference set A, refB – reference set B, refC – reference set C, refD – reference set D, pure – only descriptors.

In general it can be concluded that binary QSAR performs quite well in this instance. SIBAR methods have shown better performance than the descriptors alone though results have been not as good as in previously published studies concerning ABCB1.

4.1.2 Support vector machine

As previously mentioned a support vector machine more or less represents a black box where a separating hyperplane is developed that enables classification of the data. Nevertheless it remains the responsibility of the user to choose the right kernel necessary for optimal data separation and to select the only two user-dependent parameters in the most suitable way to render this specific classification method successful for the particular target. Though

the support vector machine generally is a robust method able to digest noise and redundant information two parameters are important for optimal performance. The regularisation parameter (or complexity constant) C and the kernel dependent variable which is defined by γ as the width in a radial basis function kernel and the exponent σ of the polynomial kernel. According to the results of Bruce and colleagues[157] a generally applicable set of parameters could not be found so that parameter optimisation by the user is still necessary.

The first study using support vector machine has been done with the package WEKA[165] using the SMO (Sequential Minimal Optimisation) method implemented in Version 3.5.7. In order to find the right parameters for the task the GridSearch function implemented in WEKA was employed and also the effects of two different kernels on this diverse datasets have been examined. Regarding the radial basis function Kernel the search for the optimal width γ ranged from 10^{-4} to 10^{5} with one step iteration. The right value for the regularisation parameter was looked for between the value of 1 to 301 with a 25 step iteration. This has been repeated for every training set with different descriptors and the respective optimal values used. Regarding the polynomial kernel the exponent σ ranged from 1 (linear kernel) to 4 with one step iteration (Table 4.5, 4.6, 4.7).

Interestingly the exponent of the polynomial kernel did not seem as susceptible to variation as the RBF kernel. Especially for the 3D Autocorrelation and the VSA descriptors a linear polynomial kernel seemed mostly sufficient for model generation. In case of the 2D descriptors a linear kernel could not succeed so that the optimal exponent was chosen to be between 3 and 4. This could be due to the type of descriptors used as the 2D descriptors mostly contain counting devices such as number of hydrogens, number of acceptable bonds the establishment of a separating hyperplane based on these type of descriptors demands another approach with a higher coefficient. On the other hand VSA and 3D Autocorrelation descriptors by their very nature encode outside features of the molecule which may complement the effect of

the support vector machine approach.

The width of the radial basis function kernel was not quite as stable as the exponent of the polynomial kernel but swayed between 0,1 and 1. Another matter entirely was the regularisation parameter which did not show any stable values but displayed a whole spectrum between the two outer borders of 1 and 301. These results underscore that although a possible generally applicable kernel parameter could be used for the datasets the regularisation parameter C remains utterly unpredictable and has to be established anew for every dataset where this specific classification method is used.

SMO-Polykernel	10CV A	A1	A0	Pr1	Pr0	C	Exponent σ
3DAuto refA	71,53	66,07	76,40	71,36	71,67	176	1
3DAuto refB	76,03	74,49	77,40	74,58	77,32	301	1
3DAuto refC	72,33	71,12	73,40	70,41	74,07	226	1
3DAuto refD	73,91	77,46	70,75	70,15	77,96	51	1
only 3DAuto	76,49	76,48	76,50	74,28	78,56	301	1
VSA refA	76,61	75,11	78,00	75,85	77,30	51	1
VSA refB	76,09	75,00	77,10	75,08	77,02	51	1
VSA refC	75,05	74,13	75,90	73,89	76,13	276	4
VSA refD	77,43	75,83	78,88	76,36	78,39	301	1
only VSA	80,53	78,33	82,50	80,11	80,88	1	3
2D refA	75,78	71,96	79,30	76,18	75,45	76	4
2D refB	79,79	83,26	76,60	76,60	83,26	301	2
2D refC	74,95	76,74	73,30	72,56	77,40	276	4
2D refD	79,28	78,61	79,88	77,85	80,58	251	3
only 2D	83,42	85,00	82,00	80,95	85,86	51	1

Table 4.5: The complexity constant C and the kernel dependent variable exponent σ used with the support vector machine approach in WEKA. refA – reference set A, refB – reference set B, refC – reference set C, refD – reference set D, 3DAuto - 3D Autocorrelation descriptors, VSA - VSA descriptors, 2D - 2D/ADME descriptors, only – pure descriptors.

SMO-RBF-Kernel	10CV A	A1	A0	Pr1	Pr0	C	γ
3DAuto refA	72,49	73,48	67,87	69,72	71,77	51	10
3DAuto refB	73,76	49,50	70,40	69,98	77,88	126	1
3DAuto refC	72,75	72,58	72,90	70,45	74,92	76	10
3DAuto refD	74,24	77,89	71,00	70,45	78,34	201	1
only 3DAuto	80,33	77,32	83,00	80,15	80,48	301	0,01
VSA refA	74,38	71,52	77,00	74,10	74,61	101	0,1
VSA refB	74,22	73,48	74,90	72,92	75,43	101	1
VSA refC	74,32	70,11	78,20	0,75	0,74	151	0,1
VSA refD	79,28	80,00	78,63	77,11	81,37	301	0,1
only VSA	80,92	82,36	79,63	78,44	83,38	301	0,1
2D refA	78,18	80,87	75,70	75,38	81,14	51	10
2D refB	74,22	76,63	72,00	71,57	77,01	126	1
2D refC	73,49	75,65	71,50	70,95	76,14	76	10
2D refD	76,18	72,64	79,38	76,02	76,32	201	0,1
only 2D	82,43	82,36	82,50	80,90	83,86	301	1

Table 4.6: The optimal complexity constant C and the radial basis function width used with the support vector machine approach in WEKA. refA – reference set A, refB – reference set B, refC – reference set C, refD – reference set D, 3DAuto - 3D Autocorrelation descriptors, VSA - VSA descriptors, 2D - 2D/ADME descriptors, only – pure descriptors.

e1071 SVM refA	C	γ	refB	C	γ
ColorScore	10	0,01	ColorScore	1	0,01
ComboScore	1	0,1	ComboScore	10	0,01
Overlap	100	0,001	Overlap	10	0,01
ScaledColor	10	0,01	ScaledColor	100	0,1
ShapeTanimoto	10	0,1	ShapeTanimoto	1	0,1
Tversky(d)	10	0,01	Tversky(d)	10	0,1
Tversky(q)	100	0,01	Tversky(q)	10	0,01
vSURF	1	0,1	vSURF	100	0,1
VSA	100	0,01	VSA	1	0,1

e1071 SVM refC	C	γ	refD	C	γ
ColorScore	10	0,1	ColorScore	1	0,1
ComboScore	1	0,1	ComboScore	100	0,001
Overlap	10	0,1	Overlap	10	0,1
ScaledColor	10	0,1	ScaledColor	1	0,1
ShapeTanimoto	10	0,1	ShapeTanimoto	10	0,1
Tversky(d)	1	0,1	Tversky(d)	10	0,1
Tversky(q)	10	0,1	Tversky(q)	10	0,1
vSURF	10	0,1	vSURF	1	0,1
VSA	100	0,01	VSA	100	0,01
only vSURF	100	0,01			
only VSA	10	0,01			

Table 4.7: The complexity constant C and the radial basis function width employed in the grid search of the support vector machine approach in R. refA – reference set A, refB – reference set B, refC – reference set C, refD – reference set D; vSURF – VolSurf descriptors, only – pure descriptors.

4.1.2.1 Radial basis function kernel

Regarding the performance of the two kernels in prediction accuracy the RBF kernel had a slight edge on the polynomial kernel. The best model achieved with the radial basis function kernel displayed an overall accuracy of 77% using SIBAR VSA descriptors with reference set B and also had an accuracy on substrates of 61% and accuracy on non-substrates of 87% and precision of 73% and 79% for substrates and non-substrates. Though in this classification approach the pure descriptors without the enhancement of the SIBAR approach also showed a creditable performance with an overall accuracy of 75%, a good accuracy on substrates of 72%, an accuracy on non-substrates of 77% and precision of 65% and 82% on substrates and non-substrates. The descriptors responsible were the VSA descriptors and also the 2D descriptors performed very well.

Still the prediction of substrates of ABCB1 seems to remain a veritable challenge as also the robust support vector machine has problems establishing a good model in this respect. Although it has to be mentioned that the accuracy on substrates works far better with this method than with the binary QSAR method. On the whole the models achieved with the RBF kernel mostly ranged between 68% to 75% (Table 4.8).[213]

Descriptors	A	A1	A0	Pr1	Pr0
3DAuto refA	56,25	61,11	53,33	44,00	69,57
VSA refA	68,75	44,44	83,33	61,54	71,43
2D refA	70,83	83,33	63,33	57,69	86,36
3DAuto refB	68,75	77,78	63,33	56,00	82,61
VSA refB	**77,08**	**61,11**	**86,67**	**73,33**	**78,79**
2D refB	64,58	72,22	60,00	52,00	78,26
3DAuto refC	72,92	83,33	66,67	60,00	86,96
VSA refC	70,83	50,00	83,33	64,29	73,53
2D refC	68,75	72,22	66,67	56,52	80,00
3DAuto refD	70,83	44,44	86,67	66,67	72,22
VSA refD	70,83	44,44	86,67	66,67	72,22
2D refD	68,75	66,67	70,00	57,14	77,78
only 3DAuto	60,42	61,11	60,00	47,83	72,00
only VSA	**75,00**	**72,22**	**76,67**	**65,00**	**82,14**
only 2D	**75,00**	**72,22**	**76,67**	**65,00**	**82,14**

Table 4.8: Performance of support vector machine on basis of radial basis function kernel. Best results are highlighted in bold letters. A – overall accuracy, A1 – accuracy on substrates, A0 – accuracy on non-substrates, Pr1 – precision on substrates, Pr0 – precision on non-substrates, 3D Auto – 3D Autocorrelation descriptors, 2D – 2D/ADME descriptors, only – pure descriptors

4.1.2.2 Polynomial kernel

The polynomial kernel though comparable cross-validated models could be presented did not perform as well. The overall accuracy of all the resulting models was found between 62% to 70% with the best model being achieved with SIBAR 2D descriptors with an overall accuracy of 71%, 78% and 66% accuracy on substrates and non-substrates and a precision of 58% and 83% for substrates and non-substrates. Surprisingly the pure 2D descriptors also achieved an overall accuracy of 71%. Generally it can be said that the best performing descriptors in this instance have been the VSA and the 2D descriptors leaving the 3D Autocorrelation descriptors behind (Table 4.9).[213]

4.1.2.3 Support vector machine in R with radial basis function kernel

Since the results of the first study using the support vector machine have been encouraging the second study also was done with the help of the support vector machine. In this approach the software R[159] was used and the support vector machine in the package e1071 employed. Again a gridSearch function was implemented and the best parameters chosen. As the RBF kernel outperformed the polynomial kernel approach only the RBF kernel was used. The resulting models again using the ROCS shape parameters as descriptors together with the VolSurf, VSA and 10 easily calculable ADME descriptors showed comparable results.

Interestingly the VolSurf SIBAR descriptors could not compare to the VSA SIBAR and ROCS SIBAR descriptors. Results of the ROCS parameters were mixed ranging from 60 % to 72% overall accuracy. The best model was achieved using reference set B together with the SIBAR Overlap and Tversky(q) parameter with an overall accuracy of nearly 72%, 62,5% accuracy on substrates, 77% accuracy on non-substrates, 59% and 79% precision on substrates and non-substrates and a Matthews Coefficient of 0.39. The best result using the pure descriptors were given by the VSA descriptors with an overall accuracy of nearly 70%, an accuracy on substrates of 62,5% and

Descriptor	A	A1	A0	Pr1	Pr0
3DAuto refA	56,25	55,56	56,67	43,48	68,00
VSA refA	70,83	55,56	80,00	62,50	75,00
2D refA	68,75	61,11	73,33	57,89	75,86
3DAuto refB	68,75	83,33	60,00	55,56	85,71
VSA refB	70,83	61,11	76,67	61,11	76,67
2D refB	**70,83**	**77,78**	**66,67**	**58,33**	**83,33**
3DAuto refC	60,42	77,78	50,00	48,28	78,95
VSA refC	68,75	50,00	80,00	60,00	72,73
2D refC	66,67	66,67	66,67	54,55	76,92
3DAuto refD	62,50	16,67	90,00	50,00	64,29
3DVSA refD	62,50	16,67	90,00	50,00	64,29
2D refD	70,83	27,78	96,67	83,33	69,05
only 3DAuto	66,67	55,56	73,33	55,56	73,33
only VSA	66,67	66,67	66,67	54,55	76,92
only 2D	**70,83**	**77,78**	**66,67**	**58,33**	**83,33**

Table 4.9: Performance of support vector machine based on the polynomial kernel. Best results are highlighted in bold letters. A – overall accuracy, A1 – accuracy on substrates, A0 – accuracy on non-substrates, Pr1 – precision on substrates, Pr0 – precision on non-substrates, refA – reference set A, refB – reference set B, refC – reference set C, refD – reference set D, 3DAuto – 3D Autocorrelation descriptors, 2D – 2D/ADME, only – pure descriptors.

an accuracy on non-substrates of 73,33% with a precision of 56% and 79% respectively. MCC was down to 0,35. Though reference set C combined with ColorScore or ScaledColor parameter received a better MCC of 0.45 and 0.53 accuracy on substrates was down to 37.50% which is very low for a predictive model (Table 4.10).

Table 4.10: Results of support vector machine approach of study 2

Testset	46 Comp.					
Support Vector Machine RBF Kernel						
	Descriptors	A	A1	A0	Pr1	Pr0
refA	ColorScore	58,70	62,50	56,67	43,48	73,91
	ComboScore	65,22	12,50	93,33	50,00	66,67
	Overlap	69,57	68,75	70,00	55,00	80,77
	ScaledColor	58,70	62,50	56,67	43,48	73,91
	ShapeTanimoto	65,22	37,50	80,00	50,00	70,59
	Tversky(d)	58,70	31,25	73,33	38,46	66,67
	Tversky(q)	67,39	56,25	73,33	52,94	75,86
	vSURF refA	63,04	87,50	50,00	48,28	88,24
	VSA refA	65,22	6,25	96,67	50,00	65,91
refB	ColorScore	69,57	43,75	83,33	58,33	73,53
	ComboScore	56,52	37,50	66,67	37,50	66,67
	Overlap	**71,74**	**62,50**	**76,67**	**58,82**	**79,31**
	ScaledColor	67,39	56,25	73,33	52,94	75,86
	ShapeTanimoto	65,22	37,50	80,00	50,00	70,59
	Tversky(d)	65,22	56,25	70,00	50,00	75,00
	Tversky(q)	**71,74**	**62,50**	**76,67**	**58,82**	**79,31**
	vSURF refB	45,65	56,25	40,00	33,33	63,16
	VSA refB	69,57	31,25	90,00	62,50	71,05
	10ADME refB	58,70	43,75	66,67	41,18	68,97
refC	**ColorScore**	**76,09**	**37,50**	**96,67**	**85,71**	**74,36**

Table 4.10: Results of support vector machine approach of study 2

Testset	46 Comp.					
Support Vector Machine RBF Kernel						
	Descriptors	A	A1	A0	Pr1	Pr0
	ComboScore	58,70	12,50	83,33	28,57	64,10
	Overlap	56,52	31,25	70,00	35,71	65,63
	ScaledColor	**78,26**	**37,50**	**100,00**	**100,00**	**75,00**
	ShapeTanimoto	63,04	37,50	76,67	46,15	69,70
	Tversky(d)	67,39	56,25	73,33	52,94	75,86
	Tversky(q)	58,70	31,25	73,33	38,46	66,67
	vSURF refC	58,70	62,50	56,67	43,48	73,91
	VSA refC	71,74	25,00	96,67	80,00	70,73
refD	ColorScore	73,91	43,75	90,00	70,00	75,00
	ComboScore	60,87	68,75	56,67	45,83	77,27
	Overlap	63,04	62,50	63,33	47,62	76,00
	ScaledColor	71,74	37,50	90,00	66,67	72,97
	ShapeTanimoto	65,22	50,00	73,33	50,00	73,33
	Tversky(d)	65,22	56,25	70,00	50,00	75,00
	Tversky(q)	60,87	62,50	60,00	45,45	75,00
	vSURF refD	63,04	75,00	56,67	48,00	80,95
	VSA refD	67,39	37,50	83,33	54,55	71,43
	10ADME refD	58,70	31,25	73,33	38,46	66,67
	only vSURF	56,52	62,50	53,33	41,67	72,73
	only VSA	**69,57**	**62,50**	**73,33**	**55,56**	**78,57**
	only 10ADME	58,70	43,75	66,67	41,18	68,97

Table 4.10: Results of support vector machine approach of study 2

Testset	46 Comp.					
Support Vector Machine RBF Kernel						
	Descriptors	A	A1	A0	Pr1	Pr0

Table 4.10: Results of support vector machine approach. Best results are highlighted in bold letters. A – overall accuracy, A1 – accuracy on substrates, A0 – accuracy on non-substrates, Pr1 – precision on substrates, Pr0 – precision on non-substrates, refA -reference set A, refB – reference set B, refC – reference set C, refD – reference set D, only – pure descriptors, vSURF – VolSurf descriptors.

Comparing the shape based descriptors of ROCS to the performance of VolSurf descriptors shape based methods outperformed the other 3D descriptors. The most effective descriptors in this regard were the ColorScore, ScaledColor, sometimes Overlap and the Tversky(q) parameter. Interestingly the VSA descriptors performed quite well which may be due to their carrying also 3D information in spite of being 2D descriptor. Generally speaking the support vector approach showed a creditable performance though the diversity of the dataset as well as the promiscuity of the protein did not allow better prediction accuracy. Regarding the overall performance of the two support vector machine approaches the performances were comparable and satisfying. Once again the prediction of substrates remains the crux of the matter. Though significantly better in this aspect than the binary QSAR approach there is still room for improvement. Although other models achieved an overall accuracy of 78% accuracy on substrates was down to 37% which is not acceptable for reliable, usable model for ABCB1 classification.[260]

The two software tools available have each their own merit and produce thoroughly comparable results even if for further research the R project is better usable. Though the first contact is hard and involves dire manual reading the software has many applicational possibilities and can be much more widely used than the WEKA package.

Recently Sato and coworkers published a study also based on support vector machine and ROCS parameters as descriptors.[261] Taking inhibitors for various target proteins from a medicinal chemistry database the shape overlay of each molecule in a compound database and all known active compounds was performed via ROCS. The resulting parameter table was taken as descriptors and a support vector machine approach performed. The training set consisted of 50 active compounds and 4950 decoys and was validated on a test set with identical properties (50 active compounds and 4950 decoys for each of the 15 target proteins). Conformational sampling was done with OMEGA[248] and a 3D similarity profile put together. For this purpose a compound from the target database was taken as query structure for ROCS and the respective similarity parameters by overlay onto the training and test set calculated. Each of the 50 active compounds of the training set was screened by ROCS and the average values calculated.

By arraying these values a 3D similarity comparable to the study here was generated and depicted the descriptors for support vector machine studies. Also in this case the ScaledColor parameter obtained the best predictive values followed by ComboScore, ColorTanimoto and ShapeTanimoto in that order. The best model could be obtained by using a combination of ScaledColor and ShapeTanimoto as input variables opposed to single variables. This result highlights the importance of pharmacophoric feature representation in combination with shape based approaches and justifies again the approach used in this study.

4.1.3 Random forest

A random forest is built on basis of an ensemble of decision trees thereby combining the interpretability of a decision tree with the better prediction performance of an ensemble method. In this study three models of random forest have been built. Two models are based on a random forest from the shelf built of 500 trees as suggested by Sventik and colleagues.[187] As randomness plays an important role in this method these two random forests were built from the same random seeds in order to explore the variation of

prediction accuracies due to different random descriptors selected for tree building.

The aspect that makes random forest such a powerful tool may present a problem when faced with the reproducibility of the method. As expected the same descriptors produced roughly the same trends in prediction accuracy although prediction accuracies differed up to 6% in the mean of overall accuracy, accuracy on substrates, accuracy on non-substrates, precision on substrates and non-substrates. Thus in order to receive reliable models for substrate prediction more than one model should be built using random forest as better models may evolve in the course of model building.

The best model so far resulting from random forest had a Matthews Coefficient of 0.43 with an overall accuracy of nearly 83%, an accuracy on substrates of 63%, an accuracy on non-substrates of 93% and further 83% each precision on substrates and non-substrates. Though this is still some way off the 88% overall accuracy achieved by Huang and colleagues[85] it has to be borne in mind that no special feature selection method was employed The model based on SIBAR VSA descriptors from reference set D was built with the entire set of descriptors in place.[260] It may be worth a try to employ feature selection algorithms before calculating the SIBAR matrix as redundant information in this way could be eliminated.

On the whole the general results were once again varied ranging from 60 to 75% overall accuracy though accuracy on substrates was markedly improved regarding also the results of the support vector machine. The accuracy on substrates ranged around 60 up to 81% with a few outliers. Regarding the performance of the reference sets reference set B of the satellite structures and also reference set D performed best. The 3D shape descriptors on the whole performed satisfactorily. The best performing parameters gained out of the ROCS shape similarity matrix were the ColorScore, the ScaledColor and the Tversky(q) parameter (Table 4.11, 4.12).

The other 3D descriptors VolSurf did not perform as well as the shape based methods although performance was better than with the support vec-

tor machine approach. Once again the VSA descriptors performed reliably and together with SIBAR produced the best model. Surprisingly with random forest the pure descriptors improved creditably with a performance of 74% overall accuracy of the VSA descriptors alone. The VolSurf descriptors performed nearly equally well with an overall prediction accuracy of 72%. The second try for a random forest model confirmed the overall trend observed with the original models but the high prediction accuracy of nearly 83 % could not be reproduced though a higher Matthews coefficient could be reached. Still the same descriptors result in the best models but with varying external prediction accuracy.

The VSA SIBAR descriptors using reference set D in this case had an overall accuracy of 80%, an accuracy on substrates of 56% and an accuracy of non-substrates of 93%. The precision on substrates was around 80% as was the precision on non-substrates with a Matthews Coefficient of 0.55. The pure VolSurf descriptors fell down to 50% overall prediction accuracy though accuracy on substrates was up to 87%.

These findings underscore the importance of model validation and the building of more than one model in search for reliable classification methods especially in combination with random based machine learning approaches like random forest.

Table 4.11: Results of first random approach a

Testset	46 Comp. Random Forest Descriptors	A	A1	A0	Pr1	Pr0
refA	ColorScore	69,57	62,50	73,33	55,56	78,57
	ComboScore	52,17	43,75	56,67	35,00	65,38
	Overlap	65,22	75,00	60,00	50,00	81,82
	ScaledColor	**71,74**	**62,50**	**76,67**	**58,82**	**79,31**
	ShapeTanimoto	54,35	37,50	63,33	35,29	65,52
	Tversky(d)	63,04	62,50	63,33	47,62	76,00
	Tversky(q)	67,39	75,00	63,33	52,17	82,61

Table 4.11: Results of first random approach a

Testset	46 Comp. Random Forest Descriptors	A	A1	A0	Pr1	Pr0
	vSURF refA	58,70	62,50	56,67	43,48	73,91
	VSA refA	60,87	0,00	93,33	0,00	63,64
refB	ColorScore	67,39	68,75	66,67	52,38	80,00
	ComboScore	60,87	62,50	60,00	45,45	75,00
	Overlap	63,04	68,75	60,00	47,83	78,26
	ScaledColor	67,39	68,75	66,67	52,38	80,00
	ShapeTanimoto	63,04	37,50	76,67	46,15	69,70
	Tversky(d)	60,87	50,00	66,67	44,44	71,43
	Tversky(q)	**71,74**	**81,25**	**66,67**	**56,52**	**86,96**
	vSURF refB	60,87	68,75	56,67	45,83	77,27
	VSA refB	60,87	12,50	86,67	33,33	65,00
	10ADME refB	63,04	31,25	80,00	45,45	68,57
refC	ColorScore	60,87	56,25	63,33	45,00	73,08
	ComboScore	58,70	56,25	60,00	42,86	72,00
	Overlap	58,70	75,00	50,00	44,44	78,95
	ScaledColor	65,22	68,75	63,33	50,00	79,17
	ShapeTanimoto	60,87	18,75	83,33	37,50	65,79
	Tversky(d)	60,87	56,25	63,33	45,00	73,08
	Tversky(q)	63,04	68,75	60,00	47,83	78,26
	vSURF refC	58,70	68,75	53,33	44,00	76,19
	VSA refC	60,87	0,00	93,33	0,00	63,64
refD	ColorScore	60,87	43,75	70,00	43,75	70,00
	ComboScore	65,22	75,00	60,00	50,00	81,82
	Overlap	58,70	56,25	60,00	42,86	72,00
	ScaledColor	65,22	50,00	73,33	50,00	73,33

Table 4.11: Results of first random approach a

Testset	46 Comp. Random Forest Descriptors	A	A1	A0	Pr1	Pr0
	ShapeTanimoto	60,87	37,50	73,33	42,86	68,75
	Tversky(d)	65,22	37,50	80,00	50,00	70,59
	Tversky(q)	65,22	62,50	66,67	50,00	76,92
	vSURF refD	69,57	81,25	63,33	54,17	86,36
	VSA refD	**82,61**	**62,50**	**93,33**	**83,33**	**82,35**
	10ADME refD	63,04	31,25	80,00	45,45	68,57
pure	only vSURF	71,74	68,75	73,33	57,89	81,48
	only VSA	**73,91**	**62,50**	**80,00**	**62,50**	**80,00**
	only 10ADME	54,35	25,00	70,00	30,77	63,64

Table 4.11: Results derived from random forest approach a. Bold letters indicate the best results. A – overall accuracy, A1 – accuracy on substrates, A0 – accuracy on non-substrates, Pr1 – precision on substrates, Pr0 – precision on non-substrates, refA -reference set A, refB – reference set B, refC – reference set C, refD – reference set D, vSURF – VolSurf descriptors, only – pure descriptors.

Table 4.12: Results from random forest approach b

Testset RF new	46 Comp. Descriptors	A	A1	A0	Pr1	Pr0
refA	ColorScore	69,57	62,50	73,33	55,56	78,57
	ComboScore	60,87	37,50	73,33	42,86	68,75
	Overlap	67,39	75,00	63,33	52,17	82,61
	ScaledColor	**69,57**	**62,50**	**73,33**	**55,56**	**78,57**
	ShapeTanimoto	50,00	31,25	60,00	29,41	62,07
	Tversky(d)	60,87	62,50	60,00	45,45	75,00

Table 4.12: Results from random forest approach b

Testset RF new	46 Comp. Descriptors	A	A1	A0	Pr1	Pr0
	Tversky(q)	67,39	75,00	63,33	52,17	82,61
	vSURF refA	60,87	62,50	60,00	45,45	75,00
	VSA refA	60,87	0,00	93,33	0,00	63,64
refB	ColorScore	65,22	56,25	70,00	50,00	75,00
	ComboScore	58,70	62,50	56,67	43,48	73,91
	Overlap	63,04	62,50	63,33	47,62	76,00
	ScaledColor	**65,22**	**68,75**	**63,33**	**50,00**	**79,17**
	ShapeTanimoto	63,04	43,75	73,33	46,67	70,97
	Tversky(d)	58,70	43,75	66,67	41,18	68,97
	Tversky(q)	**63,04**	**62,50**	**63,33**	**47,62**	**76,00**
	vSURF refB	54,35	62,50	50,00	40,00	71,43
	VSA refB	60,87	12,50	86,67	33,33	65,00
	10ADME refB	63,04	31,25	80,00	45,45	68,57
refC	ColorScore	60,87	62,50	60,00	45,45	75,00
	ComboScore	60,87	56,25	63,33	45,00	73,08
	Overlap	60,87	81,25	50,00	46,43	83,33
	ScaledColor	58,70	62,50	56,67	43,48	73,91
	ShapeTanimoto	60,87	31,25	76,67	41,67	67,65
	Tversky(d)	58,70	56,25	60,00	42,86	72,00
	Tversky(q)	63,04	75,00	56,67	48,00	80,95
	vSURF refC	60,87	68,75	56,67	45,83	77,27
	VSA refC	60,87	0,00	93,33	0,00	63,64
refD	ColorScore	63,04	50,00	70,00	47,06	72,41
	ComboScore	60,87	68,75	56,67	45,83	77,27
	Overlap	65,22	62,50	66,67	50,00	76,92

Table 4.12: Results from random forest approach b

Testset RF new	46 Comp. Descriptors	A	A1	A0	Pr1	Pr0
	ScaledColor	60,87	43,75	70,00	43,75	70,00
	ShapeTanimoto	63,04	43,75	73,33	46,67	70,97
	Tversky(d)	69,57	50,00	80,00	57,14	75,00
	Tversky(q)	60,87	50,00	66,67	44,44	71,43
	vSURF refD	65,22	75,00	60,00	50,00	81,82
	VSA refD	**80,43**	**56,25**	**93,33**	**81,82**	**80,00**
	10ADME refD	63,04	31,25	80,00	45,45	68,57
vSURF	only vSURF	50,00	87,50	30,00	40,00	81,82
VSA	**only VSA**	**73,91**	**62,50**	**80,00**	**62,50**	**80,00**
10ADME	only 10ADME	54,35	25,00	70,00	30,77	63,64

Table 4.12: Results derived from random forest approach b. Bold letters indicate the best results. A – overall accuracy, A1 – accuracy on substrates, A0 – accuracy on non-substrates, Pr1 – precision on substrates, Pr0 – precision on non-substrates, refA -reference set A, refB – reference set B, refC – reference set C, refD – reference set D, vSURF – VolSurf descriptors, only – pure descriptors.

Another aspect beside the randomness behind the forest are two parameters that can influence model performance. These regard the number of trees used for model building and the number of descriptors randomly selected at each node. For classification the default value of *mtry* is set to half the number of descriptors. *Mtry* represents the pool of descriptors from which the descriptors are randomly selected. In order to also observe accuracy variation regarding the number of trees in the forest the number was raised to 1000 trees with the best *mtry* selected with the aid of the training set. The best *mtry* value was taken and a regular model built with random forest.

Though Breiman[183] already postulated that the forest is relatively immune to changes in *mtry* the tuning algorithm as installed in the software package R was put to use. Surprisingly nearly the opposite result could be observed. Instead of tuning the forest and thereby enhancing the model performance overall accuracies sometimes decreased. In any case no substantial change in effect regarding overall accuracy could be observed. This may be due to the afore-mentioned injected randomness of random forest and so sheer luck decides whether the upcoming model will be better than the one previously built. Another reason for the non-effect of tuning the forest may be the high number of trees that are built with no pruning in random forest. Maybe though real overfitting should not be possible in such a forest the branches and nodes became too bushy for real classification purposes. It could also be true that the *mtry* default used by the regular random forest method as implemented in the R package really is the optimal value for all classification purposes and represents an omnipotent parameter for binary classification problems.

As best classification model once again was built on VSA SIBAR descriptors based on reference set D with an overall accuracy of 76% though accuracy on substrates was down to 44%. The Matthews correlation coefficient for these SIBAR descriptors amounted to 0.45. Another very good model could be built using the Overlap parameter of reference set A with an accuracy on substrates of 75%, an overall accuracy of 67% and an acceptable accuracy on non-substrates of 63%. Once again the pure descriptors performed quite well indeed even outperforming the SIBAR based descriptors as the pure VSA descriptors were responsible for an overall accuracy of 78% an accuracy on substrates of 63%, an accuracy on non-substrates of 87% and 71% and 81% precision on substrates and non-substrates. The Matthews correlation coefficient for this model scored 0.51. Though performing satisfactorily before the pure VolSurf descriptors were down to 52% overall prediction accuracy which once again renders the VSA descriptors clear winner in the duel. Again the most significant descriptors of the shape based approach were the ROCS derived ColorScore, Overlap and the Tversky(q) parameter. (Table 4.13).

Table 4.13: Results from tuned random forest

Testset 46 Comp. RF tuned	Descriptors	A	A1	A0	Pr1	Pr0
refA	ColorScore	67,39	62,50	70,00	52,63	77,78
	ComboScore	52,17	37,50	60,00	33,33	64,29
	Overlap	**67,39**	**75,00**	**63,33**	**52,17**	**82,61**
	ScaledColor	67,39	56,25	73,33	52,94	75,86
	ShapeTanimoto	54,35	37,50	63,33	35,29	65,52
	Tversky(d)	60,87	62,50	60,00	45,45	75,00
	Tversky(q)	63,04	75,00	56,67	48,00	80,95
	vSURF refA	58,70	68,75	53,33	44,00	76,19
	VSA refA	60,87	0,00	93,33	0,00	63,64
refB	ColorScore	63,04	56,25	66,67	47,37	74,07
	ComboScore	63,04	62,50	63,33	47,62	76,00
	Overlap	63,04	62,50	63,33	47,62	76,00
	ScaledColor	60,87	56,25	63,33	45,00	73,08
	ShapeTanimoto	**65,22**	**56,25**	**70,00**	**50,00**	**75,00**
	Tversky(d)	63,04	50,00	70,00	47,06	72,41
	Tversky(q)	65,22	62,50	66,67	50,00	76,92
	vSURF refB	54,35	56,25	53,33	39,13	69,57
	VSA refB	60,87	12,50	86,67	33,33	65,00
	10ADME refB	60,87	31,25	76,67	41,67	67,65
refC	ColorScore	63,04	56,25	66,67	47,37	74,07
	ComboScore	65,22	62,50	66,67	50,00	76,92
	Overlap	52,17	75,00	40,00	40,00	75,00
	ScaledColor	60,87	62,50	60,00	45,45	75,00
	ShapeTanimoto	60,87	25,00	80,00	40,00	66,67
	Tversky(d)	58,70	56,25	60,00	42,86	72,00

Table 4.13: Results from tuned random forest

Testset	46 Comp.					
RF tuned						
	Descriptors	A	A1	A0	Pr1	Pr0
	Tversky(q)	60,87	68,75	56,67	45,83	77,27
	vSURF refC	56,52	62,50	53,33	41,67	72,73
	VSA refC	60,87	0,00	93,33	0,00	63,64
refD	ColorScore	60,87	50,00	66,67	44,44	71,43
	ComboScore	65,22	68,75	63,33	50,00	79,17
	Overlap	60,87	50,00	66,67	44,44	71,43
	ScaledColor	63,04	50,00	70,00	47,06	72,41
	ShapeTanimoto	58,70	37,50	70,00	40,00	67,74
	Tversky(d)	60,87	37,50	73,33	42,86	68,75
	Tversky(q)	63,04	56,25	66,67	47,37	74,07
	vSURF refD	65,22	75,00	60,00	50,00	81,82
	VSA refD	**76,09**	**43,75**	**93,33**	**77,78**	**75,68**
	10ADME refD	67,39	37,50	83,33	54,55	71,43
pure	only vSURF	52,17	87,50	33,33	41,18	83,33
	only VSA	**78,26**	**62,50**	**86,67**	**71,43**	**81,25**
	only 10ADME	52,17	18,75	70,00	25,00	61,76

Table 4.13: Results derived from tuned random forest. Bold letters indicate the best results. A – overall accuracy, A1 – accuracy on substrates, A0 – accuracy on non-substrates, Pr1 – precision on substrates, Pr0 – precision on non-substrates, refA -reference set A, refB – reference set B, refC – reference set C, refD – reference set D, vSURF – VolSurf descriptors, only – pure descriptors.

4.1.4 Further classification approaches

4.1.4.1 Linear discriminant analysis

By a visiting professor the possible merit of linear discriminant analysis for this special problem was pointed out. Taking the advice the linear discriminant analysis was carried out using once again the R project with the package lda. Linear discriminant analysis is similar to principal component analysis with a distinct difference. Principal component analysis regards the importance and possible information contained in the variables whereas discriminant analysis lays its focus on the differentiation of the two classes not regarding the variables as priority. It is defined by a prior possibility (random) of one compound belonging to either of the classes and a posterior probability after model generation. It is more or less the precursor of the support vector machine and also follows Bayes theorem. This analysis is simple and easily interpretable though collinearity and noisy data may be a problem and feature selection is often necessary. Nevertheless in order not to bias the results and to stay comparable no feature selection was performed.

The results did not provide many insights as the prediction accuracy was much lower than with the other methods and sometimes no model could be built. The best model achieved with linear discriminant analysis was surprisingly provided by VolSurf SIBAR descriptors based on reference set D with an overall accuracy of nearly 72% but with an accuracy on substrates of 38% and 90% on non-substrates. Though overall accuracy is not bad an accuracy on substrates with less than 50 percent is not acceptable for any reliable model. Again the accuracy on substrates was the stepping stone of the analysis and the overall results ranged from 55% to 60% overall accuracy.

The Tversky(q), the ScaledColor and the Overlap descriptors sometimes produced an error message and no model could be established. The VSA descriptors showed satisfactory overall accuracy though the accuracy on substrates was disastrous and sometimes down to 0% which means that all the compounds in the training set have simply been classified as non-substrates thereby effecting an accuracy on non-substrates of 100%. The pure descriptors did not perform favourably in comparison with the SIBAR descrip-

tors with the best approach of the VSA descriptors achieving an overall accuracy of nearly 61%, an accuracy on substrates of 38%, an accuracy on non-substrates of 73% and 43% and 69% precision on substrates and non-substrates.

In conclusion, though useful and highly informative when working, in this particular case the linear discriminant analysis did not provide any new insights or highly predictive models (Table 4.14).

Testset	46 Comp. LDA Descriptors	A	A1	A0	Pr1	Pr0
refD	vSURF	71,74	37,50	90,00	66,67	72,97
refD	VSA	65,22	43,75	76,67	50,00	71,88
refD	10ADME	67,39	56,25	73,33	52,94	75,86
	only VSA	60,87	37,50	73,33	42,86	68,75

Table 4.14: Depiction of the best models produced by LDA. A – overall accuracy, A1 – accuracy on substrates, A0 – accuracy on non-substrates, Pr1 – precision on substrates, Pr0 – precision on non-substrates, refA -reference set A, refB – reference set B, refC – reference set C, refD – reference set D, vSURF – VolSurf descriptors, only – pure descriptors.

4.1.4.2 Quadratic discriminant analysis

In the cases where linear discriminant analysis does not work quadratic discriminant analysis is suggested as another option. For this reason quadratic discriminant analysis again as implemented in the lda package of the R project was employed. More or less similar results could be observed. Though overall accuracies did not look too bad accuracy on substrates was in many cases down to 0% which does not suggest any predictive abilities of the model. The accuracies of this aspect ranged form 0% to 25 which is in no way acceptable for any model. One of the best models regarding overall accuracy could be achieved by the ComboScore based on reference set D with an overall accuracy of 74%, an accuracy on substrates of 44%, an accuracy

on non-substrates of 90%, a precision on substrates of 70% and a precision on non-substrates of 75%. In this particular case the VolSurf descriptors outperformed the VSA descriptors but with very bad overall results indicating the failure of the method on this particular dataset, these particular descriptors and the particular target. (Table 4.15)

Testset QDA	46 Comp. Descriptors	Acc	A1	A0	Pr1	Pr0
refA	ColorScore	69,57	12,50	100,00	100,00	68,18
refA	ComboScore	67,39	6,25	100,00	100,00	66,67
refA	ScaledColor	69,57	12,50	100,00	100,00	68,18
refB	10ADME refB	67,39	6,25	100,00	100,00	66,67
refD	ComboScore	73,91	43,75	90,00	70,00	75,00
refD	vSURF refD	67,39	6,25	100,00	100,00	66,67
refD	VSA refD	67,39	6,25	100,00	100,00	66,67

Table 4.15: Depiction of the best models produced by QDA. A – overall accuracy, A1 – accuracy on substrates, A0 – accuracy on non-substrates, Pr1 – precision on substrates, Pr0 – precision on non-substrates, refA -reference set A, refB – reference set B, refC – reference set C, refD – reference set D, vSURF – VolSurf descriptors, only – pure descriptors.

4.1.5 Comparison methods used

In conclusion it can be stated that the random forest though not as reproducible as sometimes would be hoped outperformed binary QSAR and support vector machine with the former method following swiftly. Support vector machine in this particular assignment did not perform as well as the binary QSAR method. The challenge for all the classification algorithms always represented the identification of substrates of ABCB1 whereas non-substrates were easier to classify. This implicates a problem as the utmost desire of pharmaceutical industry is the reliable identification of ABCB1 substrates because these compounds cause many problems in clinical trials especially for cancer patients. On the other hand reliable prediction of non-substrates is also important as ABCB1 has moved from possible saving angel

via inhibition in cancer therapy to untouchable anti-target.

In summary regarding all the methods especially for this promiscuous target and the highly diverse natural product database higher complex machine learning methods have to be employed to lead to success. Though easier interpretable methods naturally are desirable the most important thing in every modelling approach is the predictive performance of the model. Thus it is also highly undesirable to interpret models that have no predictive power because that in itself renders interpretation futile. Therefore random forest would be method of choice in any future modelling approach. Though support vector machine is known for its robustness in this study binary QSAR showed the higher applicability with the added advantage of implemented descriptor importance measurements as well as the random forest. The support vector machine alone works as a black box where the derivation of the separating hyperplane cannot be shown.[260]

4.2 Descriptors

4.2.1 2D versus VSA versus 3D Autocorrelation descriptors

As explained earlier the first study in this work has been comprised of three different descriptor sets namely a set of simple 2D descriptors, a set of orthogonal van der Waals Surface Area area based descriptors and 3D based Autocorrelation descriptors. Regarding the performance of these descriptors results again have been mixed. The best model achieved with 3D Auocorrelation vectors yielded an overall accuracy of 73% with an accuracy on substrates of 83%, an accuracy on non-substrates of nearly 67% a precision of 60% on substrates and 87% on non-substrates whereas the best model based on 2D descriptors had an overall accuracy of 75% with an accuracy on substrates of 56% on substrates and 87% on non-substrates and a precision of 71% versus 77% on non-substrates. The best performance of VSA descriptors has been

given with an accuracy on 77%, an accuracy on substrates of 61% and an accuracy on non-substrates of 87% with a precision of 73% on substrates and 79% on non-substrates (Table 4.16).

Descriptor	A	A1	A0	Pr1	Pr0	Method
3DAuto refC	72,92	83,33	66,67	60,00	86,96	SVM-RBF
VSA refB	77,08	61,11	86,67	73,33	78,79	SVM-RBF
2D refA	75,00	55,56	86,67	71,43	76,47	BQSAR -15 C.
only 3DAuto	66,67	55,56	73,33	55,56	73,33	SVM-Polyk
only VSA	75,00	72,22	76,67	65,00	82,14	SVM-RBF
only 2D	75,00	72,22	76,67	65,00	82,14	SVM-RBF

Table 4.16: Best models of every descriptor set. A – overall accuracy, A1 – accuracy on substrates, A0 – accuracy on non-substrates, Pr1 – precision on substrates, Pr0 – precision on non-substrates, only – pure descriptors, 3DAuto – 3D Autocorrelation descriptors, 2D – 2D/ADME descriptors, SVM – support vector machine, RBF – radial basis function kernel, Polyk – polynomial kernel, BQSAR 15C – binary QSAR restrained to 15 components.

In an overall comparison of all the models generated the VSA descriptors yielded the best general performance though the 2D descriptors descriptors did almost equally well. The worst performance however has been shown by the 3D Autocorrelation descriptors with a few good prediction results but also many bad outliers This result seems to highlight a fact Labute himself stated namely that VSA descriptors really are orthogonal descriptors and with their compactness of only 32 are also able to describe the properties of a molecule better than a set of 88 2D descriptors and 84 3D Autocorrelation descriptors. Though officially 2D descriptors VSA descriptors contain „pseudo-3D“ information in the form of van der Waal surface area paired with information on molar refractivity, log P and electronic charges that in effect emphasise the features of the molecules. Nevertheless, it has to be stated that no conformational sampling has been done prior to the calculation of the Autocorrelation descriptors thereby maybe not utilising the full potential of the 3D Autocorrelation vectors. Also as the aim was the comparison of the

applicability of shape based SIBAR methods to purely 2D or 3D calculated descriptors we refrained from any feature selection.

4.2.2 10 ADME versus 2D VSA versus 3D VolSurf versus shape based descriptors

In order to demonstrate the usability of sophisticated descriptors in combination with the SIBAR approach 10 ADME descriptors consisting of the number of hydrogen bond acceptors, hydrogen bond donors, the number of rotatable bonds, lipophilicity, molecular refractivity, negative and positive partial charges, van der Waal Surface area, weight and polar surface area have been calculated. The mean performance of only these descriptors lay beyond the performance of the other descriptors though the SIBAR approach seemed to enhance the prediction accuracy. The best model of the 10 ADME descriptors combined with SIBAR received an external accuracy of 67% using the tuned random forest or the linear discriminant analysis whereas the best performance of the pure 10 ADME descriptors received only 58,7% external accuracy (Table 4.17). This demonstrates the effect of the SIBAR approach and suitability of shape based approaches when classifying ABCB1 substrates and non-substrates especially on a widely diverse natural product dataset. However, once again the accuracy on substrates seems to be the crux of the matter as the best result provided only 56% accuracy on substrates opposed to an accuracy on non-substrates of 73% (Table 4.14).

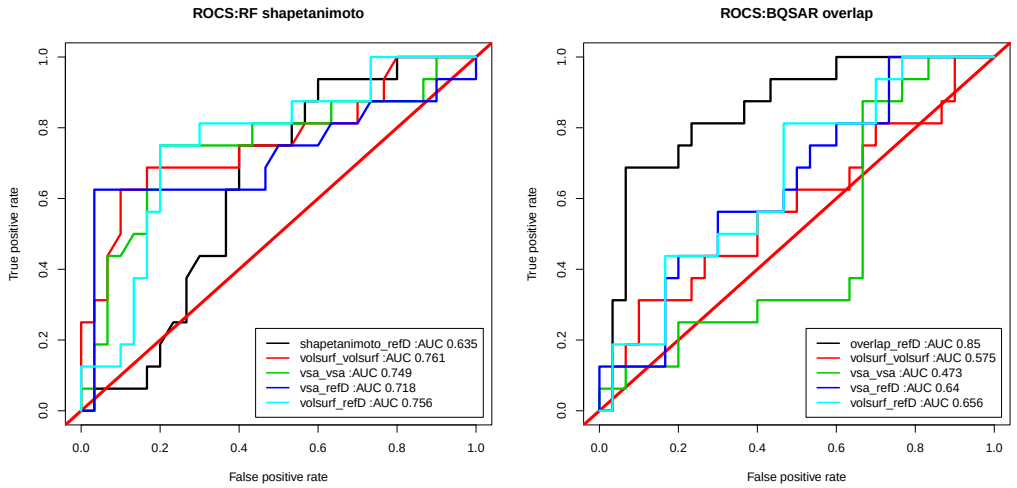
As concluded with the results of the first study the overall accuracy of VSA descriptors in general was a bit more stable than with 3D VolSurf descriptors and rendered acceptable results. In fact the best model of this study has been built with VSA SIBAR descriptors using reference set D and the random forest classification method with an overall accuracy of nearly 83%, an accuracy on substrates of 63% and an accuracy on non-substrates of 93%. The precision for either substrate or non-substrate lay around 83%. Though also bad outliers occurred with the worst overall accuracy around 54% with binary QSAR the performance on the whole was okay.

Descriptor	A	A1	A0	Pr1	Pr0	Method
10ADME refD	67,39	37,50	83,33	54,55	71,43	tuned RF
only 10ADME	58,70	43,75	66,67	41,18	68,97	SVM
VSA refD	82,61	62,50	93,33	83,33	82,35	RF
only VSA	73,91	62,50	80,00	62,50	80,00	RF new
vSURF refD	69,57	81,25	63,33	54,17	86,36	RF
only vSURF	71,74	68,75	73,33	57,89	81,48	RF

Table 4.17: Best models of the descriptor sets. A – accuracy, A1 – accuracy on substrates, A0 – accuracy on non-substrates, Pr1 – precision on substrates, Pr0 – precision on non-substrates, refD- reference set D, vSURF – VolSurf descriptors, only – pure descriptors.

Observing the results with respect to the reference sets it can be concluded that the performance of VSA descriptors may vary due to different reference sets. With the reference sets derived from the ChemGPS approach performance was worse than with the tailored reference set. Also it has to be said that VSA descriptors as 2D descriptors are very easy to calculate, have the benefit of containing also 3D information of the van der Waals Surface area and are orthogonal descriptors thereby redundancy should not be a problem. Used as pure descriptors without SIBAR the VSA descriptors performed quite well with the best model built with the tuned random forest with an accuracy of 78%, an accuracy on substrates of 63%, an accuracy on non-substrates of 87% and precision values of 71% and 81% respectively. Compared with other descriptors used alone without SIBAR VSA descriptors provided the best models by far. Figure 4.1a depicts the model on a receiver operator curve (ROC) that is based on the true positive against the false positive rate. The area under the curve (AUC) is another measure of model performance. Values of 1 render optimal prediction, 0.5 is random prediction and lower than random means no significant model could be built. The straight red line shows random prediction.

The 3D VolSurf descriptors contain information on size and shape, hy-



(a) random forest (b) binary QSAR

Figure 4.1: (a) random forest model based on reference set D, (b) binary QSAR model based on reference set D. ShapeTanimoto_refD – ShapeTanimoto parameter based on reference set D, overlap_refD - Overlap parameter based on reference set D, refD – SIBAR reference set D, volsurf_volsurf – pure VolSurf descriptors, vsa_vsa – pure VSA descriptors, vsa_refD – VSA descr. with reference set D, volsurf_refD – VolSurf Descr. with reference set D.

drophilic and hydrophobic regions as well as interaction energy moments and therefore were considered a suitable vehicle for the 3D based SIBAR approach. On the whole the performance of these descriptors was creditable though not very stable. The best performance combined with SIBAR was achieved with random forest classification resulting in an accuracy of nearly 70% with an accuracy on substrates of 81%, an accuracy on non-substrates of 63% and a precision on substrates and non-substrates of approximately 54% and 86%. The best performance of pure VolSurf descriptors also was received using random forest with an accuracy of nearly 72% with an accuracy on substrates of 69%, an accuracy on non-substrates of 73% and a precision on substrates of 58% and 81% on non-substrates. The worst performance on the other hand other than with linear discriminant analysis was found with binary QSAR classification as no serious model could be built (Figure 4.1b).

The purely shape based method of using the parameters produced by ROCS (rapid overlay of chemical structures) has given once again varied results. The output exists of seven parameters with different meaning in respect to shape. First of all the ColorScore is a value of distribution of functional groups between the two molecules studied and according to the authors has some similarity to pharmacophore matching. The ScaledColor more or less represents the ColorScore but in scaled format which means the minimum lies around 0 and the maximum ScaledColor around 1. ShapeTanimoto represents the similarity index Tanimoto based on shape overlap. Other similarity measures are the two Tversky indices consisting of Tversky(d) and Tversky(q) which are also based on the overlap between the molecules though take into account pre-factors and are usually counted to calculate the similarity between fingerprints. The overlap between the two molecules represents the absolute overlap of volume between the two molecules in question and the ComboScore combines the ShapeTanimoto index with the ColorScore thereby containing purely shape driven information and information on molecular features.

The performance of these parameters used as descriptors for machine learning and classification was very varied. Interestingly the use only of shape as the ShapeTanimoto provides mostly did not give good results. Accuracies were found between 50% to 65% with also varying accuracy on substrates and non-substrates. Of all the ROCS parameters ShapeTanimoto showed the worst overall performance. The Overlap parameter provided robust results in classification generally ranging from 58 to 65% with a few outliers in each direction. The worst overall accuracy of 52% was rendered using the tuned random forest and the second best model in comparison with other descriptors was given using binary QSAR classification and reference set D with an overall accuracy of 80%.

As explained the Tversky indices also represent a method of calculating similarity between two molecules. Interestingly the performance of the two Tversky indices differed a bit. The Tversky(d) parameter provided overall

accuracies ranging from 55% to around 70% though mostly lying around 60 to 62%. The Tversky(q) index on the other side showed performances ranging from 58% to 72% though mostly lying around 65%. The best performing parameters regarding shape based descriptors have been the ColorScore, ScaledColor and the Tversky(q) parameters. The ComboScore parameter did not entirely live up to the expectations though most of the overall accuracies were respectable ranging from very low 52% with tuned random forest to 74% using binary QSAR. ColorScore and ScaledColor parameters represent two sides of the same coin as one parameter contains the scaled values of the other. Overall accuracies derived from the ColorScore lay in a range from 59% to 76% and ScaledColor values with the exception of one bad outlier (52%) moved in the zone of 58% up to 78% both with support vector machine.

Again due to very varying results conclusions are difficult to draw. Nevertheless it has been a surprise that the ComboScore which contains most of the information considering the other parameters did not bring better results. Though the bad outcome of the ShapeTanimoto as machine learning descriptor may provide an explanation. It seems that shape similarity alone is not the most important parameter or simply the Tanimoto calculation method is not suitable for this approach as the Tversky(q) index has fared much better. Therefore the ComboScore may contain noise not needed for model building and may not be the best descriptor in this regard as it simply is the summation of both values without any weighting done.

Another point of interest lies in the fact that though both Tversky indices performed quite acceptable the Tversky(q) index was a bit better. Interestingly this effect was turned around when regarding reference set D consisting of very large natural products formerly part of the dataset. For the reference sets based on the ChemGPS idea Tversky(q) was the better parameter. This may have its reason in the α pre-factor whose setting is the only difference between the two parameters. In Tversky(q) the query molecule is regarded as the main self-overlap term whereas in Tversky(d) the database molecule has the main self-overlap. The molecules contained in the satellite reference sets are very small whereas the compounds in the dataset are highly complex

natural compounds. Depending on the molecules in the reference sets the respective parameter should be chosen (Figure 4.2).

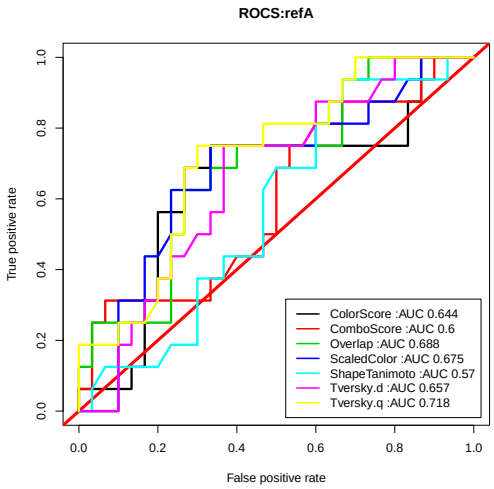


Figure 4.2: Model based on random forest and reference set A. Only parameter of ROCS are depicted.

These results are strengthened by the work of Kirchmair *et al.*.[249] who performed a study on the optimisation of shape based screening. Virtual screening was performed on the directory of useful decoys (DUD) database with conformational sampling done by the OMEGA[248] software of Openeye. Experimental structures of the ligands in the binding pocket were available for 39 targets and different approaches for the conformation of query structures applied. In direct comparison to docking considering the protein structure and purely ligand-based methods the latter achieved comparable results and therefore represent a valid approach for virtual screening.

They also investigated the performance of ROCS using the bioactive conformation as crystallised and calculated conformations of query structures as input. Their results show that no difference could be objectively detected between the bioactive conformation and even high energy conformations generated with OMEGA. Implemented in ROCS is also the possibility for multi-conformational input structures where the algorithm loops over every con-

former of the query structure and the conformers of the database compounds in order to find the best overlay. Also this method was tested and its possible benefits did not justify the much higher computational cost and no higher performance could be observed.

Their results legitimate the approach in this study to use the reference compounds as low energy conformers as query structures for ROCS. In a detailed look at the different scoring parameters provided by ROCS the group drew similar conclusions as has been done in this study. Also the ShapeTanimoto scoring alone did not meet the expectations but the ColorScore nearly performed as well as the ComboScore which represents a combination of the two. In consequence they optimised the ComboScore with higher weights assigned to the ColorScore and could enhance the screening performance though this had to be done for every dataset individually. Their study also emphasises the importance of pharmacophoric features being taken into account in shape based screening.

4.2.3 Overall comparison of descriptors

This study has been done on the prerequisite that similar shape equals similar interaction profile as the interaction between ligand and transporter is purely driven by 3D shape complementarity. For this reason 10 ADME descriptors, a set of 2D descriptors, VSA descriptors, 3D Autocorrelation vectors, 3D VolSurf descriptors and ROCS parameters have been calculated on basis of the SIBAR approach. Results show varying answers.

Comparing 2D descriptors with VSA and 3D Autocorrelation vectors the VSA descriptors showed the overall best performance whereas the 3D Autocorrelation vectors have not justified the high expectations. In the second study the VSA descriptors were compared to 3D VolSurf descriptors and ROCS parameters. The performances of these descriptors have been highly similar and VSA descriptors and ROCS parameters have provided the highest accuracies. Further analysis with a new random forest, linear discriminant and quadratic discriminant analysis have given further insights into the model prediction accuracies of the shape based method ROCS. Occasionally a se-

lection of ROCS parameters namely the ColorScore, Overlap and Tversky(q) have shown higher mean overall accuracies than VSA descriptors with whom the best model in total could be achieved.

However, the results have not differed in a significant level. This may be also explained by the high promiscuity shown by ABCB1 and as some evidence suggests the same molecule could bind at different binding sites. It follows that the site of binding may be connected with different conformations of the molecule and therefore descriptors able to quantify shape or volume in a more general way may be beneficial over purely conformation dependent approaches. Similar results have been obtained in a comparative study between 3D shape based, 2D fingerprints and docking as methods in virtual screening.[262] 2D Fingerprint similarity with the Tanimoto coefficient and atom pair similarity were calculated. Additionally Phase[101] of Schrödinger was used for 3D shape based screening. The same query structures as low energy conformations were taken for both methods. In this study 2D fingerprint methods outperformed the shape based approach, assessed on basis of the enrichment factor and the area under the curve. Though multiple query structures were taken no higher effect than with 2D methods could be observed. Compared to docking both ligand based approaches showed higher enrichment rates and higher area under the curve when assessed by ROC (receiver operating curve) curves proving once again the high merit of ligand based approaches.

ROCS is a good beginning as it uses conformational sampling only to find the best possible overlap between two molecules and further calculates the respective shape parameters. Though actually 2D descriptors VSA descriptors represent 3D property and render a more general description of the molecules. With respect to computational costs 3D based methods generally are more expensive and time consuming than 2D-based ones and regarding the small differences in significance between the descriptors used VSA descriptors seem to represent a plausible alternative if time pressure is existent. They seem to encompass enough 3D information to make shape similarity possible and applicable.

4.3 Reference set

Another important facet of this study has been the exploration of the usability of a generally applicable reference set for the SIBAR approach. For this reason four reference sets have been developed based on two different strategies. As explained earlier the first strategy was based on the idea of satellite structures on the fringes of the chemical space acting as sort of fixed stars for better navigation. Accordingly three different methods to derive those satellite structures have been employed and three reference sets generated.

A further approach has been conceived due to in-house results regarding a set of propafenone inhibitors[118] where a reference set of dataset tailored molecules was used and performed well. Therefore a fourth reference set of 40 molecules taken from the original training set has been compiled.

In the first study using 3D Autocorrelation vectors, 2D descriptors and VSA descriptors no clear best reference set could be identified. Nevertheless, regarding the mean overall accuracies reference set D seemed not to perform as well as the satellite structure approach regarding reference set A to C. The accuracy on substrates seemed to be the most difficult aspect. For the other reference sets again no obvious best set could be identified though reference set B seemed to show the most stable performance and especially worked well with the VSA descriptors.

In order not to bias the results of this study the satellite reference sets have been derived for every descriptor anew. That means that in total nine different reference sets have been employed in the first study. Results show however that no significant difference between performances could be detected. It follows that the compounds selected for each reference set are extreme enough to be used as global reference sets.

Therefore for the second study only the three satellite sets derived from VSA descriptors have been put to use to act as globally applicable reference set.

Using the shape based approach with ROCS, 3D VolSurf descriptors and VSA descriptors a difference in performance between tailored reference set

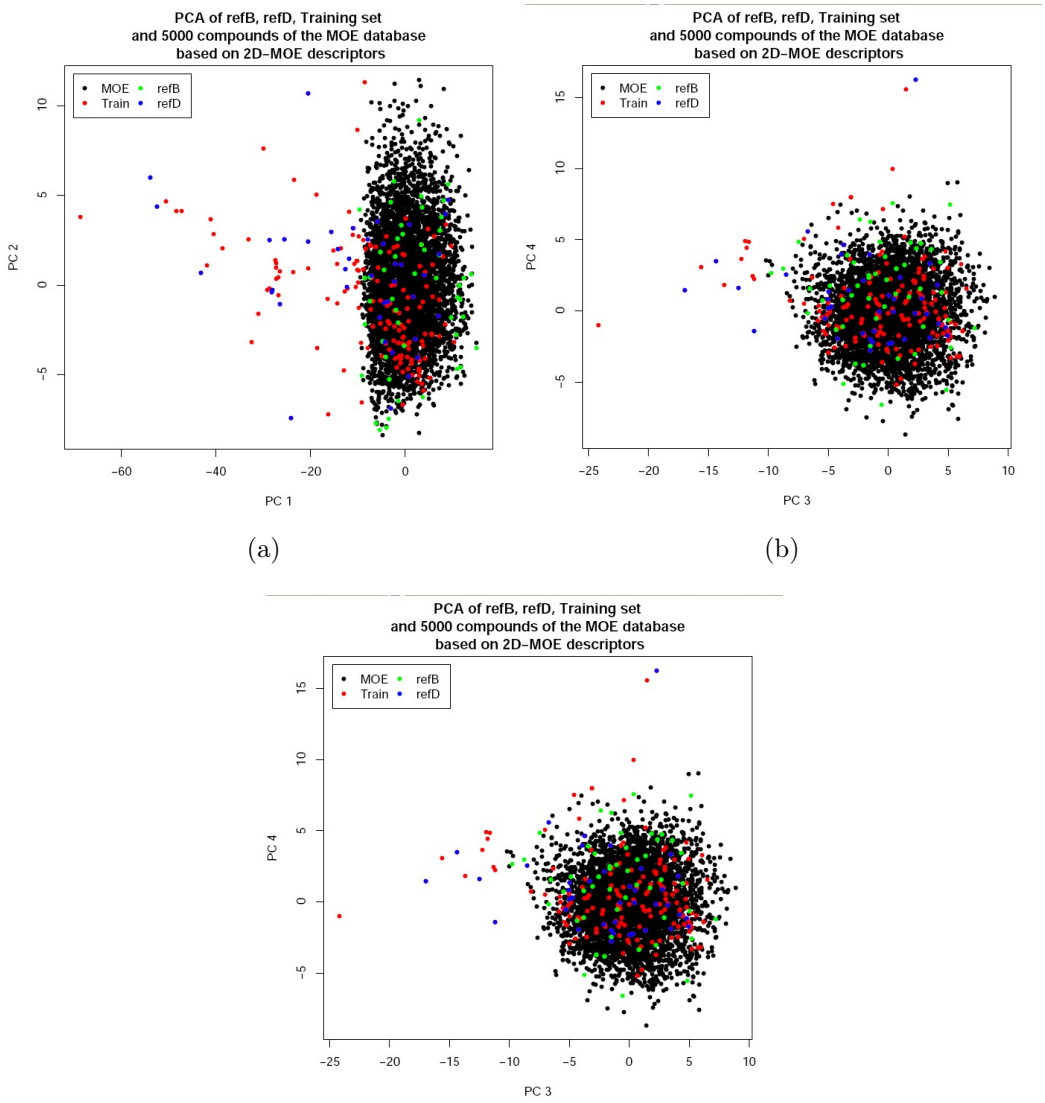
and the satellite structures could be observed. Other than expected the tailored reference set managed to outperform the previously better satellite structures though on a hair's breadth. Regarding the satellite structures the performances varied with classification method and descriptor type but overall comparison based on mean overall accuracies allows the identification of reference set A as the best performing satellite set in this case.

In order to demonstrate the diversity of the data set and the disparity between natural product compounds and „normal" chemicals Figure 4.3 illustrates the coverage of the chemical space of reference set B, reference set D, the training set and 5000 representative compounds from the MOE database. It is clearly visible that the NCI-60 compounds cover a higher amount of chemical space than the MOE representative compounds thereby illuminating their structural complexity and the singularity of these compounds. Recently Larsson and colleagues[263] stated the need for a separate ChemGPS for natural products which is emphasised by the results of this study.

Though once again no clear conclusions could be drawn the first study using the support vector machine in WEKA and the binary QSAR method of MOE suggested the higher relevance of the satellite structure approach opposed to the tailored reference set. However, in the second study the tide has turned and the tailored reference set seemed to generate higher prediction accuracies than the satellite structures. Figures 4.4 to 4.8 give an overview over the reference sets used.

This may be due to various factors: In order to prepare the dataset for the rapid shape overlay of ROCS conformational sampling had to be done using the program OMEGA. A limit for conformational sampling is represented in a limit in the number of rotatable bonds with 20 being the maximum. Therefore molecules (7) contained in the dataset with more than 20 rotatable bonds had to be excluded leaving less molecules in the training set than in the former study. This may explain the different performance of reference set D combined with different classification methods and different descriptors.

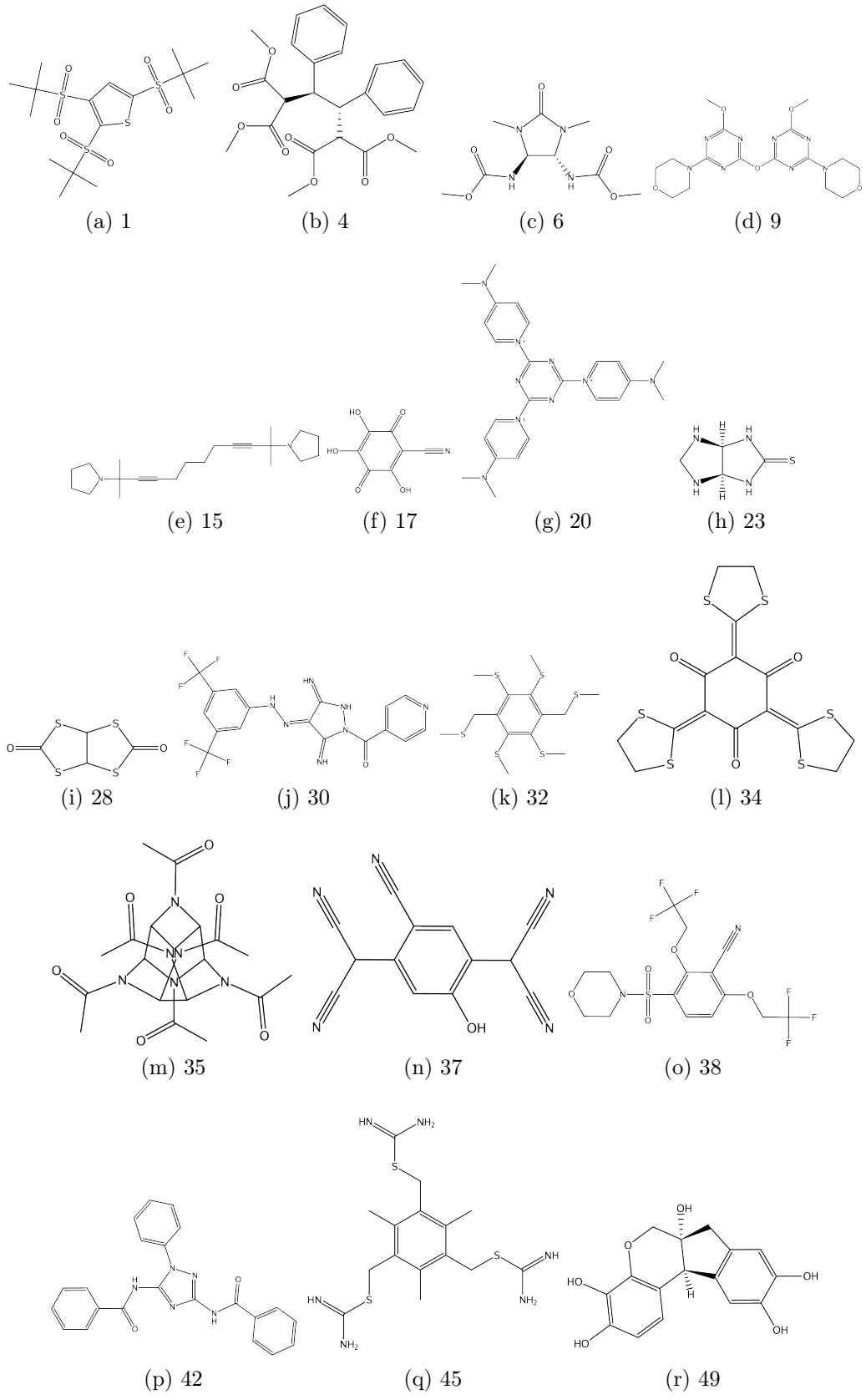
On the whole due to the small differences between the performance of the



(a) (b)

(c)

Figure 4.3: Principal component analysis(PCA) of reference sets B (refB) and D (refD), the training set and 5000 diverse compounds of the MOE database based on 2D-MOE descriptors. a) PCA of components 1 and 2 (explaining 58.8 % of the variance), b) PCA of components 2 and 3 (explaining 65% of the variance), c) PCA of components 3 and 4 (explaining 71% of the variance).



(a) 1 (b) 4 (c) 6 (d) 9

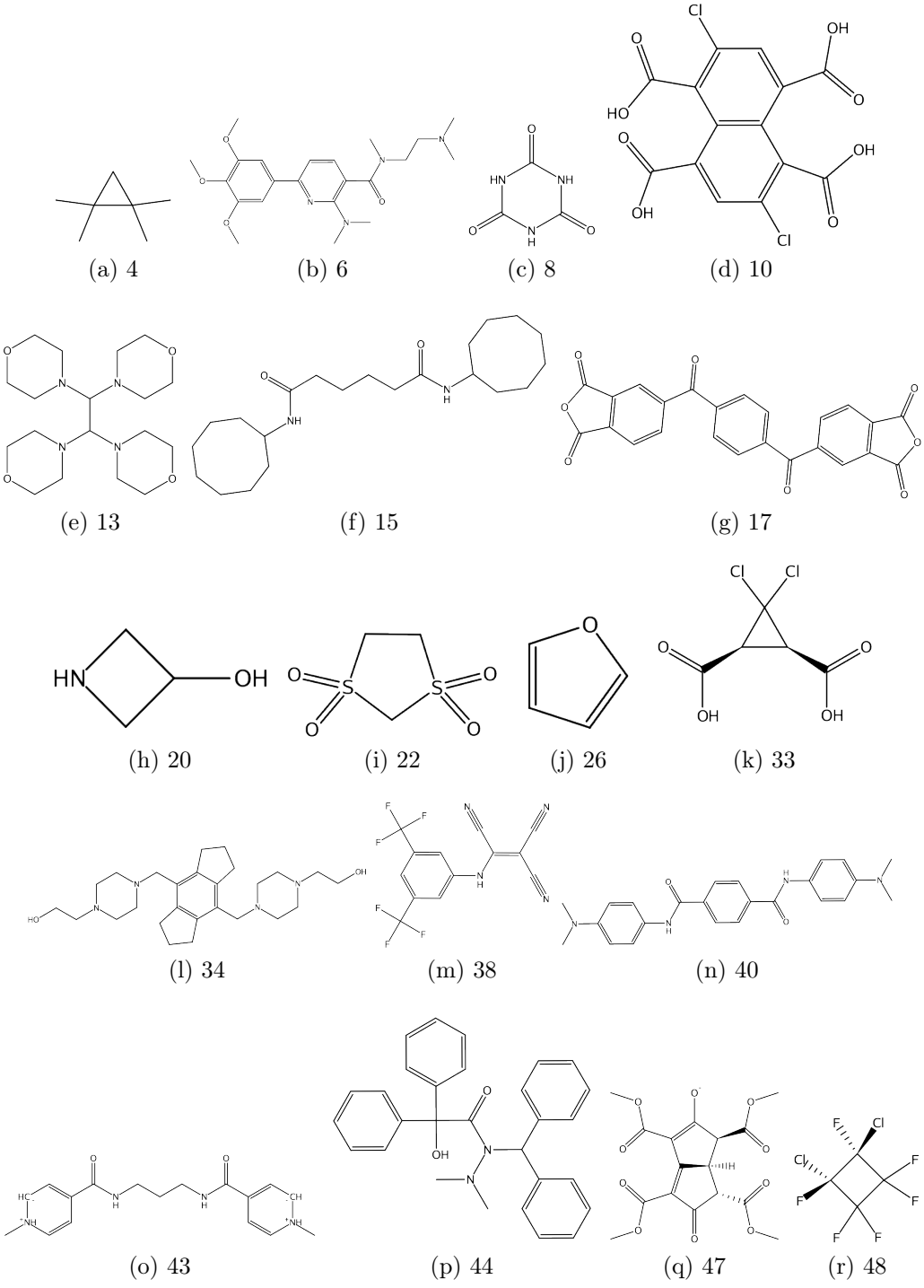
(e) 15 (f) 17 (g) 20 (h) 23

(i) 28 (j) 30 (k) 32 (l) 34

(m) 35 (n) 37 (o) 38

(p) 42 (q) 45 (r) 49

Figure 4.4: Selection of 18 compounds of reference set A.



(a) 4 (b) 6 (c) 8 (d) 10

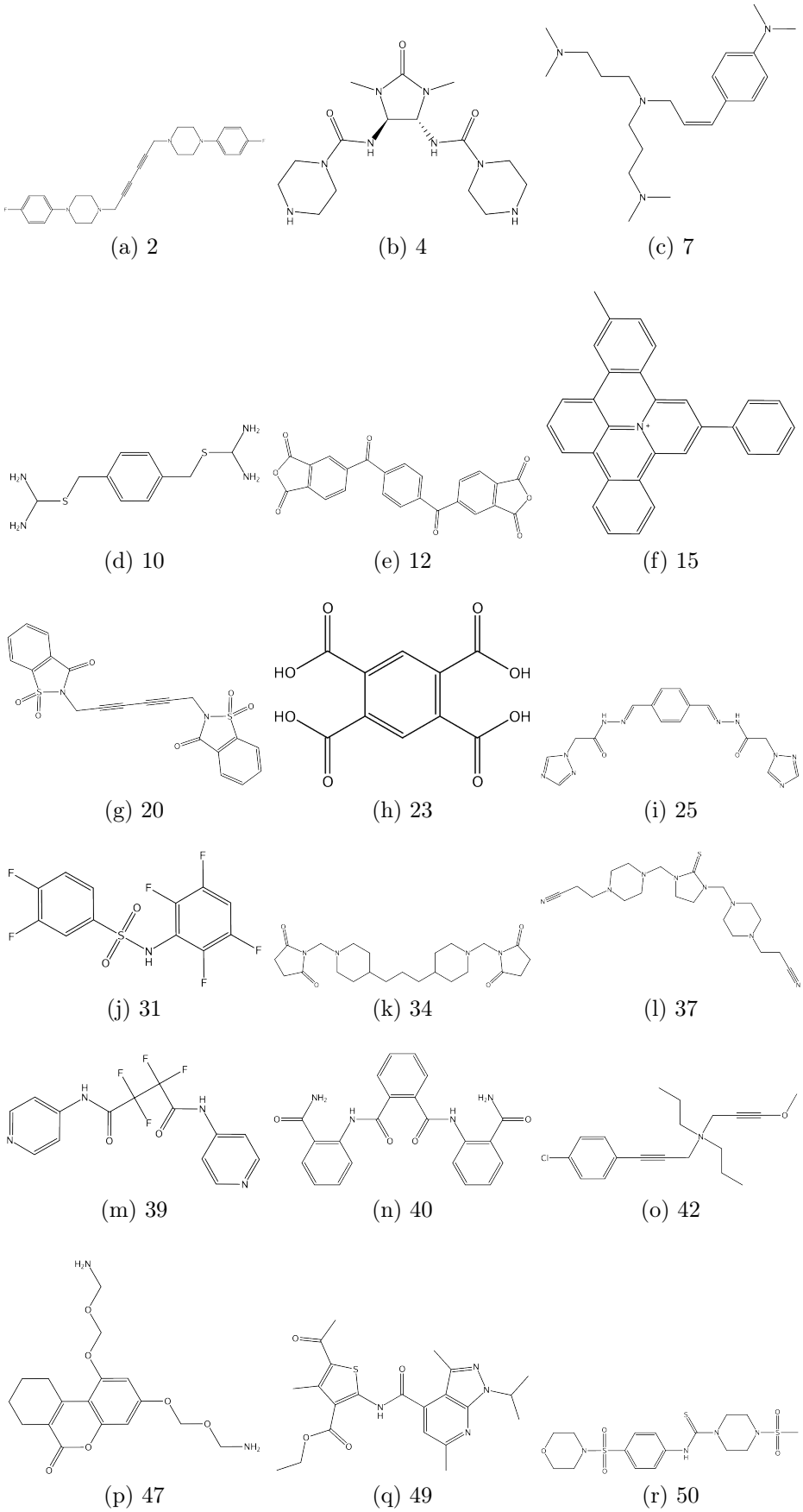
(e) 13 (f) 15 (g) 17

(h) 20 (i) 22 (j) 26 (k) 33

(l) 34 (m) 38 (n) 40

(o) 43 (p) 44 (q) 47 (r) 48

Figure 4.5: Selection of 18 compounds of reference set B.



(a) 2 (b) 4 (c) 7

(d) 10 (e) 12 (f) 15

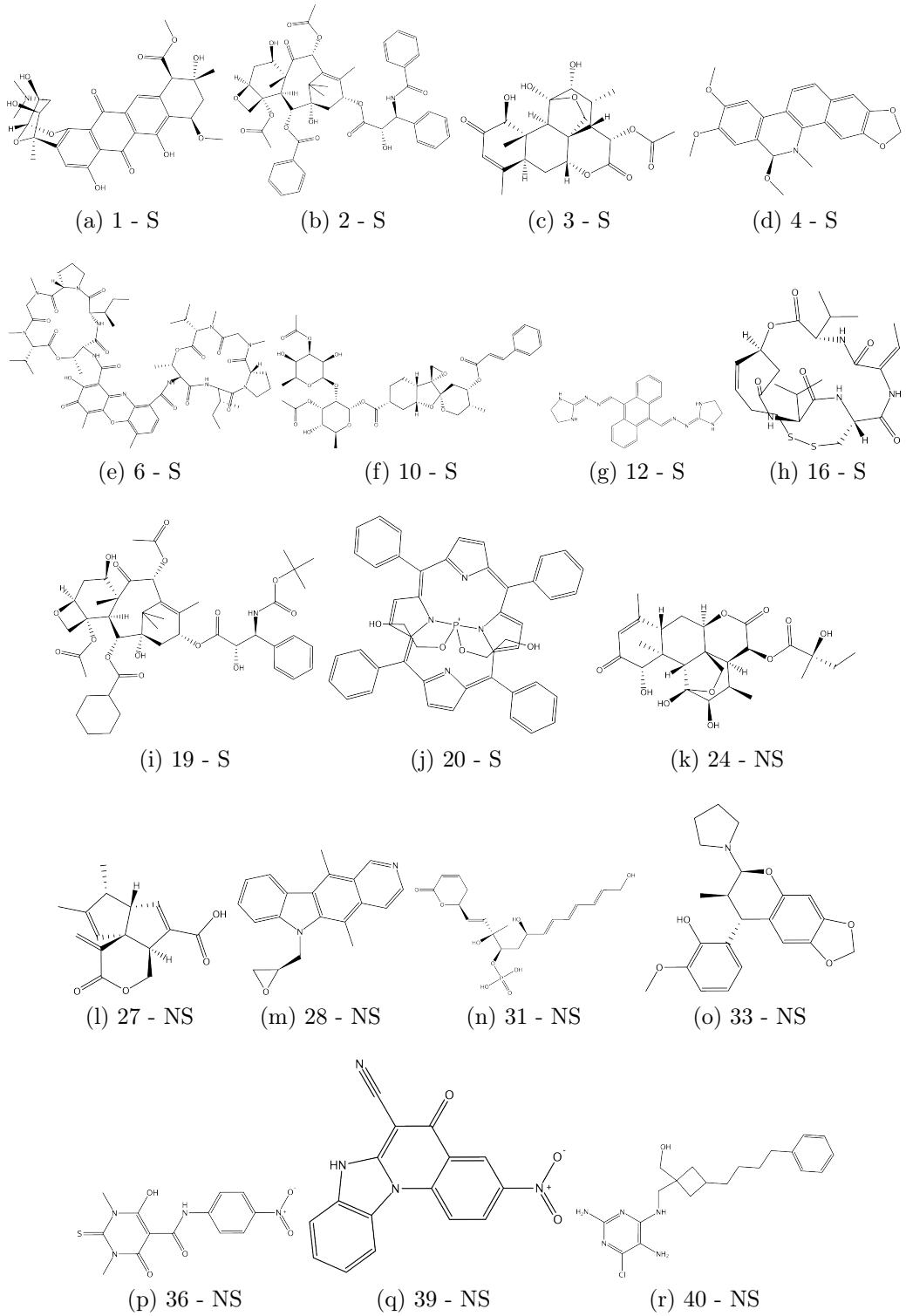
(g) 20 (h) 23 (i) 25

(j) 31 (k) 34 (l) 37

(m) 39 (n) 40 (o) 42

(p) 47 (q) 49 (r) 50

Figure 4.6: Selection of 18 compounds of reference set C.



(a) 1 - S (b) 2 - S (c) 3 - S (d) 4 - S

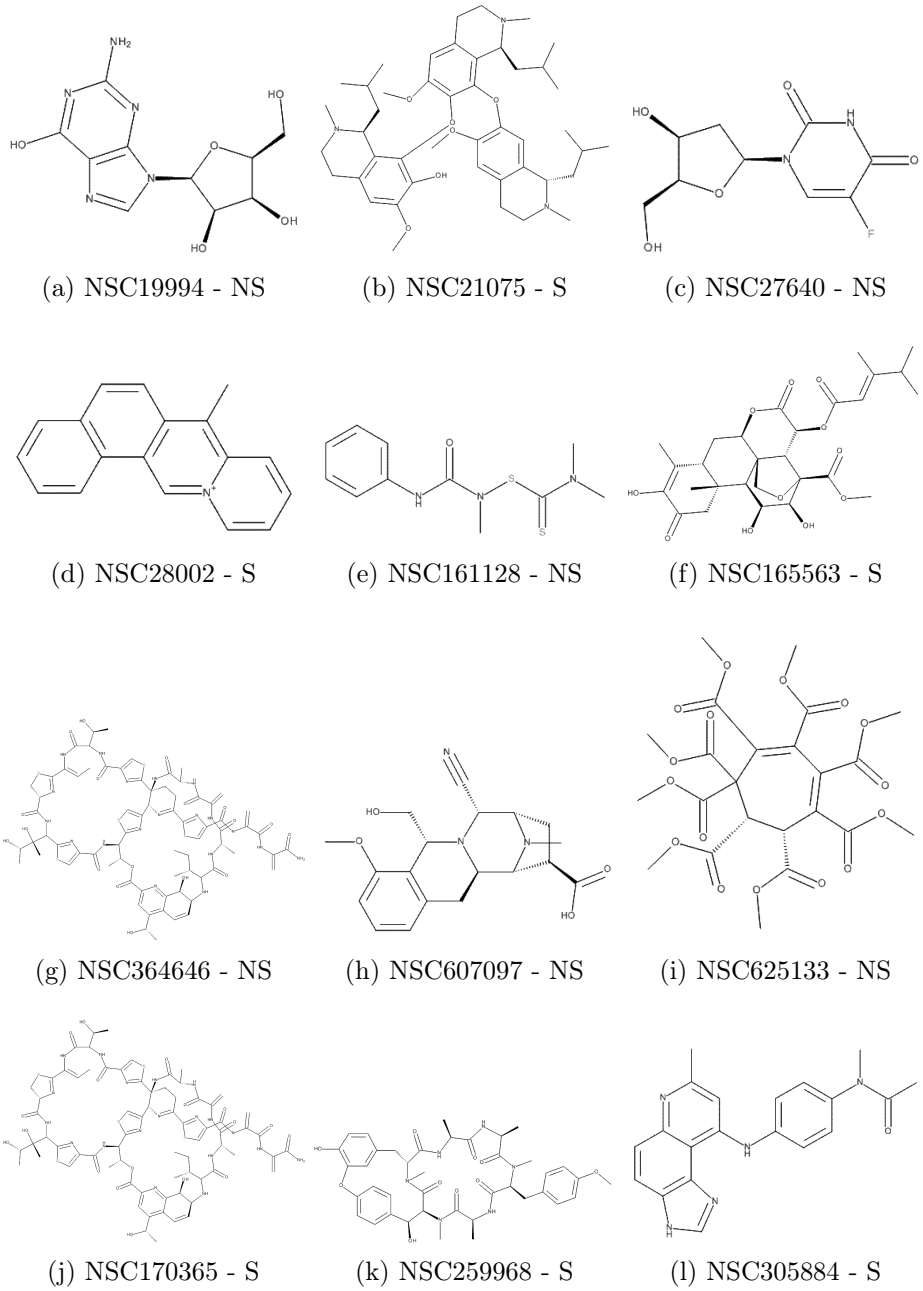
(e) 6 - S (f) 10 - S (g) 12 - S (h) 16 - S

(i) 19 - S (j) 20 - S (k) 24 - NS

(l) 27 - NS (m) 28 - NS (n) 31 - NS (o) 33 - NS

(p) 36 - NS (q) 39 - NS (r) 40 - NS

Figure 4.7: Selection of 18 compounds of reference set D. The last eight compounds are annotated as non-substrates. S- Substrate, NS – Non-substrates.



(a) NSC19994 - NS (b) NSC21075 - S (c) NSC27640 - NS

(d) NSC28002 - S (e) NSC161128 - NS (f) NSC165563 - S

(g) NSC364646 - NS (h) NSC607097 - NS (i) NSC625133 - NS

(j) NSC170365 - S (k) NSC259968 - S (l) NSC305884 - S

Figure 4.8: Selection of compounds of the dataset demonstrating its diversity and structural complexity. S- Substrate, NS – Non-substrates. Taken from Schwaha[260]

Pairwise similarity		Metric	Similarity	Distance
refA	3DAuto	Tanimoto	0,36	0,64
	2D	Tanimoto	0,22	0,78
	VSA	Tanimoto	0,28	0,72
refB	3DAuto	Tanimoto	0,34	0,66
	2D	Tanimoto	0,22	0,78
	VSA	Tanimoto	0,20	0,80
refC	3DAuto	Tanimoto	0,35	0,65
	2D	Tanimoto	0,22	0,78
	VSA	Tanimoto	0,28	0,72
refD		Tanimoto	0,39	0,61
test		Tanimoto	0,38	0,62
train		Tanimoto	0,38	0,62

Table 4.18: Pairwise similarity based on MACCS fingerprints is shown. RefA – reference set A, refB – reference set B, refC – reference set C, refD – reference set D, test – test set, train -training set, 2D – 2D/ADME descriptors, 3DAuto – 3D Autocorrelation descriptors.

individual reference sets the following course of action is proposed. In order to gain a quick overview over one's data the application of reference set A as generally applicable reference set is justified and may bring further insights. For a more intense research the selection of a tailored reference set consisting of the most diverse structures of the data set is advisable.

To demonstrate the diversity of the reference sets as well as the diversity of the NCI-60 dataset Tanimoto dissimilarity matrices have been calculated. By subtracting the computed mean Tanimoto similarity based on MACCS fingerprints within a dataset or the mean Tanimoto similarity between two datasets from 1 the mean dissimilarity has been assessed (Table 4.18). This table shows that the plan to extract satellite structures with extreme values and high dissimilarity has been successful as indeed the dissimilarity within

the satellite reference sets A to C is higher than within reference set D. It can be observed that for reference set B the Tanimoto dissimilarity is the highest confirming again the diversity among compounds.

Dissimilarity matrix			Metric	Similarity	Distance
train	refA	3D Auto	Tanimoto	0,27	0,73
		2D	Tanimoto	0,25	0,75
		VSA	Tanimoto	0,28	0,72
	refB	3DA Auto	Tanimoto	0,29	0,71
		2D	Tanimoto	0,24	0,76
		VSA	Tanimoto	0,23	0,77
	refC	3D Auto	Tanimoto	0,26	0,74
		2D	Tanimoto	0,25	0,75
		VSA	Tanimoto	0,28	0,72
	refD		Tanimoto	0,37	0,63
test	refA	3D Auto	Tanimoto	0,27	0,73
		2D	Tanimoto	0,26	0,74
		VSA	Tanimoto	0,30	0,70
	refB	3DA Auto	Tanimoto	0,30	0,70
		2D	Tanimoto	0,26	0,74
		VSA	Tanimoto	0,24	0,76
	refC	3D Auto	Tanimoto	0,27	0,73
		2D	Tanimoto	0,26	0,74
		VSA	Tanimoto	0,30	0,70
	refD		Tanimoto	0,36	0,64
train	test		Tanimoto	0,35	0,65

Table 4.19: Dissimilarity between reference sets and training and test set. RefA – reference set A, refB – reference set B, refC – reference set C, refD – reference set D, train – training set, test – test set, 2D – 2D/ADME descriptors, 3DAuto – 3D Autocorrelation descriptors.

Also the Tanimoto dissimilarity values for the training set and the test set have been calculated. The mean Tanimoto dissimilarity within each of the two datasets amounts to 0,62 again speaking of high diversity. When computing the dissimilarity between training and test set the mean Tani-

moto dissimilarity receives 0,65. This means that even if the test set has been selected from the original dataset no high similarity exists within this set. Training and test set are highly dissimilar (Table 4.19).

Another brick in the wall is the reference set D that shows less internal dissimilarity than the satellite structure reference sets but has initially been compiled out of the original NCI-60 dataset. Though for this set the most diverse compounds have been selected out of the dataset the dissimilarity within the training set and the dissimilarity within reference set D is almost the same. Also a high dissimilarity between reference set D and either training or test set has been assessed. Once again the high diversity of the natural compound database is emphasised as even by extracting the most diverse compounds from the set no higher similarity has resulted.

4.4 Variable importance

As both the random forest method and the binary QSAR provide the possibility of determining descriptor importance for these two methods descriptor importance has been measured and analysed. In this manner a number of so-called „core“ structures could be identified that seem to provide much of the information necessary for model building.

The core structures are defined as follows: For each classification method descriptor importance was ascertained and the ten most important compounds of each reference set for the respective model and descriptor set taken into consideration. Further on those compounds that were present based on at least four different descriptor sets were investigated. The core structures of each reference set have been chosen by at least four different descriptors among the top ten important structures and occurred in all two regarded classification methods.

As only fifty reference compounds and therefore fifty SIBAR descriptors are available for each dataset the selection of the same structures among the top ten most important molecules by the different descriptor models is a

matter of probability calculations. In order to avoid any chance encounters and the resulting over-interpretation these rules have been set. Chances that with both classification methods same compounds are selected as most important descriptors by at least four different descriptor sets are extremely low and point to no random but a deliberate choice. Also by combining the results of two classification methods possible biased effects of random forest descriptor selection may be prevented.

4.4.1 Reference set A

For reference set A compound 17 (Figure 4.9) was identified as core structure. Using the classification method binary QSAR it was put among the ten most important SIBAR compounds by the ColorScore, ShapeTanimoto, Tversky(d), VSA and VolSurf descriptors. With random forest it was elected into the top ten by the ColorScore, ComboScore, Overlap, ScaledColor, Tversky(q) and VSA descriptors.

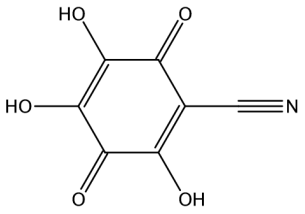
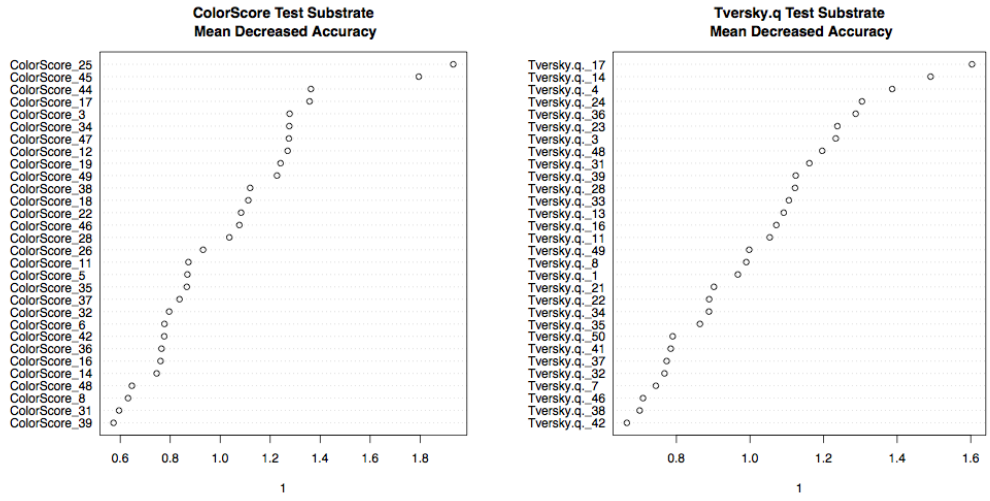


Figure 4.9: Compound 17 of reference set A,

Though being a highly unusual and chemically extreme compound it seems to hold certain key features necessary for ABCB1 substrate and non-substrate model building. Regarding the models of the descriptors deeming this compound very important some clues may be found for interpretation. Especially for binary QSAR the descriptors selecting this compound as most important showed an accuracy ranging from 54% (ShapeTanimoto) to 72% (ColorScore) but all models have in common that accuracy on non-substrates was very high whereas accuracy on substrates proved to be extremely low especially regarding the VSA descriptors that could find no

model for substrate/non-substrate classification receiving an accuracy on substrates of 0% as well with random forest.



(a) (b)

Figure 4.10: Plot of variable importance derived from random forest based on mean decreased accuracy.

As random forest allows the extraction of important descriptors for substrate prediction this compound earns a closer look. Though substrate prediction is very low with binary QSAR random forest based on the Overlap parameter reached an accuracy on substrates of over 75%. The model derived by the ColorScore descriptor achieved an accuracy on substrates of 50% and an accuracy on non-substrates of 83% with binary QSAR. An accuracy of nearly 70% was received with random forest. Accuracy on non-substrates reached 73% opposed to 62,5% accuracy on substrates. A possible conclusion may be that the core compound is important differentiating the properties of non-substrates for binary QSAR. For random forest it is important for classifying substrates of ABCB1 (Figure 4.10).

4.4.2 Reference set B

Regarding reference set B compound 22 (Figure 4.11) seems to incorporate the same importance as compound 17 for reference set A. For the method

binary QSAR this compound of reference set B was included as one of the most important compounds of the reference set by the descriptors ColorScore, ComboScore, Overlap, ScaledColor and ShapeTanimoto. Using the classification method random forest the VSA, VolSurf, ColorScore, ComboScore, ScaledColor, Overlap and Tversky(q) descriptors all enclosed compound 22 into their top ten compounds of reference set B.

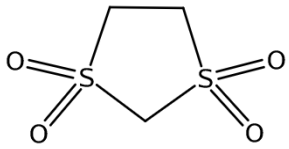
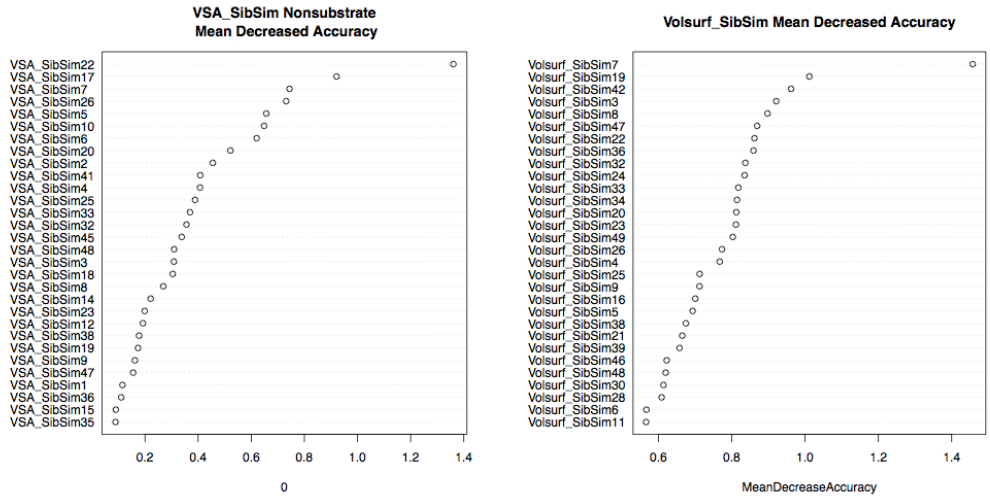


Figure 4.11: Compound 22 of reference set B.

Again this compound apparently contains some essential features that help to characterise ABCB1 substrate and non-substrate properties. Once more looking at the accuracies of the models built the overall accuracies reached from 52 to 72% using binary QSAR and from 54 to 65% using random forest. Again the accuracies on non-substrates are exceeding by far the accuracies on non-substrates but not as drastically as the former models. The models with descriptors not containing this compound in the top ten most important molecules at least with binary QSAR showed slightly worse accuracy on substrates. The accuracies on substrates ranged from 25 (binary QSAR) to 69% (random forest). Variable importance based on random forest shows the importance of this compound in prediction especially of non-substrates whereas for VolSurf descriptors this compound is essential for substrate and non-substrate prediction. This fact undermines the hypothesis that these core structures may help to identify substrate and non-substrate properties during the SIBAR approach (Figure 4.12).

4.4.3 Reference set C

For reference set C no such compound met the criteria of being voted into the top ten most important compounds by at least four descriptors and both classification methods. Though some of the molecules were chosen by both



(a) (b)

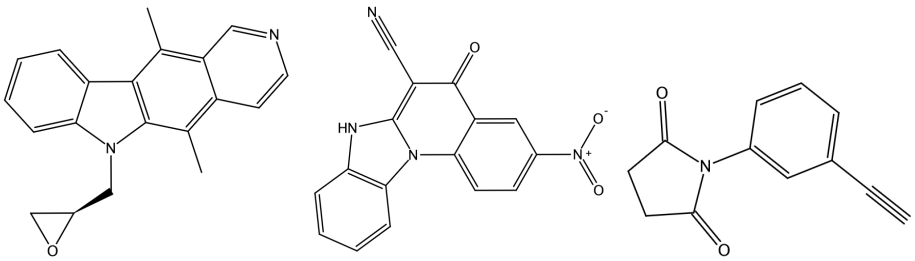
Figure 4.12: Plot of variable importance derived from random forest based on mean decreased accuracy.

classification methods as top ten members the number of descriptors fell below the rules imposed on the nomination of core structures.

4.4.4 Reference set D

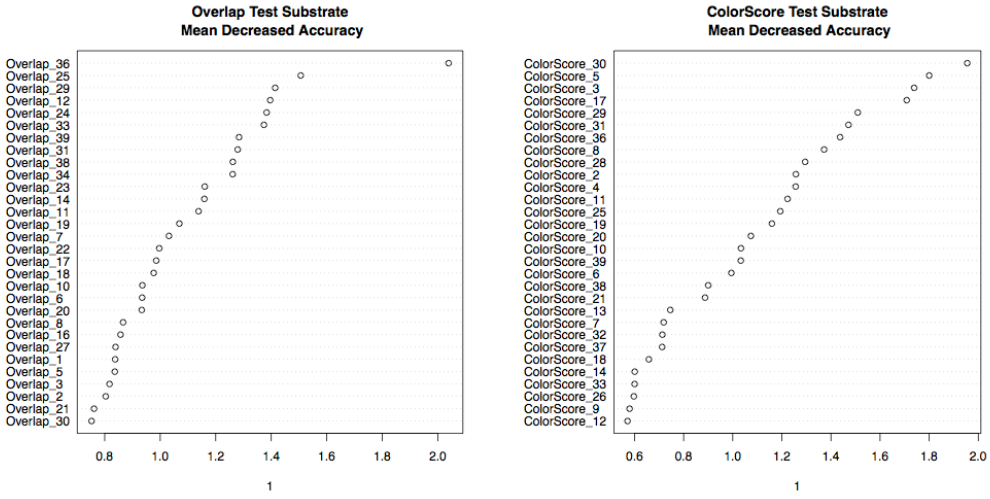
Regarding reference set D on the other side three core structures important for model building were identified. Compound 31(Figure 4.13a) was identified as important molecule of reference set D by the ColorScore, ComboScore, ShapeTanimoto, Overlap and Tversky(d) descriptors with binary QSAR as classification method. Random forest descriptors selecting this compound were the VSA, ColorScore, ComboScore, Overlap, ScaledColor, ShapeTanimoto, Tversky(q) descriptors (Figure 4.14).

Another of those core structures is compound 33 (Figure 4.13b). This molecule was put among the top ten most important compounds of reference set D by the ColorScore, ComboScore, ShapeTanimoto, Tversky(d) descriptors based on binary QSAR classification method. ComboScore, Overlap, ShapeTanimoto, VolSurf descriptors employed by the random forest machine learning method reached the same conclusion regarding this compound.



(a) Compound 31 (b) Compound 33 (c) Compound 36

Figure 4.13: Core structures of reference set D.



(a) (b)

Figure 4.14: Variable importance based on random forest and parameters Overlap and ColorScore.

The last compound of these core structures is given by compound 36 (Figure 4.13c) which has been selected with the binary QSAR method by the descriptors VolSurf, VSA, ShapeTanimoto and Tversky(d). With the random forest method the descriptors ColorScore, Overlap, ScaledColor and Tversky(q) have been nominators of this compound for the label core structure.

These three structures are annotated as non-substrates and therefore may possess special features or the placement of special features that makes binding to ABCB1 impossible. Interestingly those compounds for random forest are necessary for prediction of substrates. The prediction accuracies for the respective models with binary QSAR reach from 61% up to 80% with acceptable performance concerning the accuracy on substrates. Random forest also reached good results from 61 to 80% overall accuracy and again accuracy on substrates was acceptable from 44 up to 75%.

The identification of those core structures may provide further insights into the features important for ABCB1 substrates and non-substrates and shape similarity seems to be the right direction to go. Of course binary QSAR provides more reliable results as the descriptor importance is actually based on principal component analysis. Random Forest on the other side by its very course of action randomly puts together a set of descriptors from which the best usable descriptor for decision nodes is picked. This mechanism encompasses the whole benefit of random forest and its high predictive ability is due to the randomness in the forest on the other hand however descriptor importance can only be measured when the descriptor really is used (Figure 4.15).

By calculating the out-of-bag rate and leaving out single descriptors descriptor importance is measured. Though not utterly reliable but when combined with binary QSAR descriptor importance any bias has to disappear.

In the following section other structures will be presented that may also have great impact on model building but did not completely meet the core structure criteria. That means that for both classification methods the com-

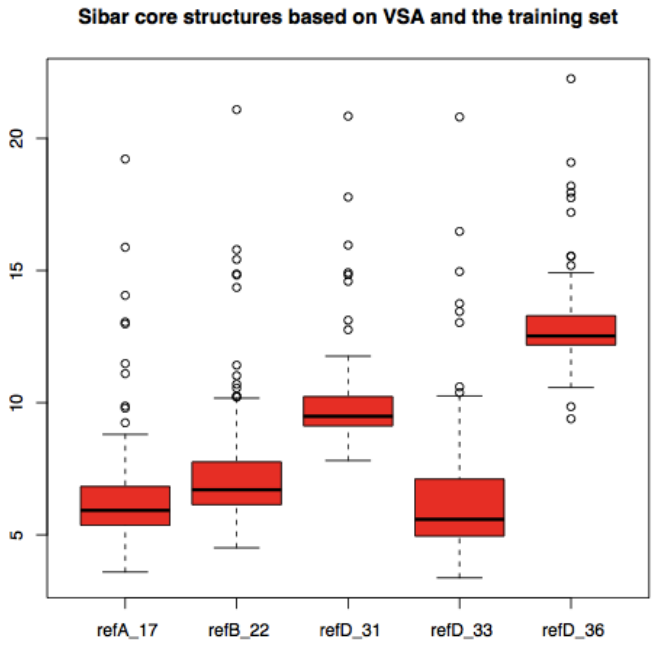


Figure 4.15: The core compounds depicted with their values in the training set based on VSA descriptors. RefA_17 – core compound reference set A, refB_22 – core compound reference set B, refD – core compounds of reference set D.

pound has been selected but not always by four descriptors of each method. Therefore the following structures could not be regarded as core structures but nevertheless may provide ample clues regarding shape similarity and ABCB1.

Accuracies of models with those descriptors and reference sets are difficult to interpret as especially for the members of the reference sets A to C no substrate or non-substrate status is known. The case lies differently with reference set D whose compounds can be annotated as either substrates or non-substrates. In order to identify further important structures for SIBAR the formerly derived descriptor importance with the first study is also added.

In this manner compound 5 could be identified as further core structure. In this case the core structure represents a substrate and may therefore concentrate special features necessary for substrate recognition in shape sim-

ilarity. The structure has been selected as very important for descriptors 3D Autocorrelation and VSA in the first study using binary QSAR and also with descriptors ColorScore and VSA in the second study. Concerning random forest this structure was chosen by descriptors ColorScore, ComboScore, Overlap and ScaledColor. (Figure 4.16)

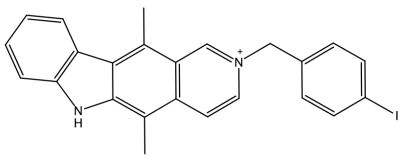


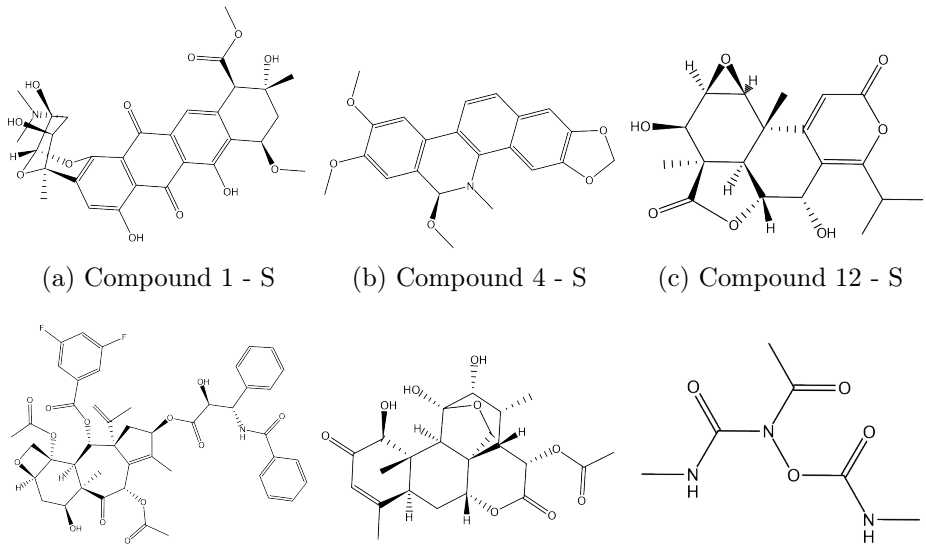
Figure 4.16: Compound 5 (substrate).

What springs to mind regarding those compounds is the fact that often the ColorScore is among the descriptors deeming those core structures important. It has to be borne in mind that the ColorScore more or less represents pharmacophoric features of the molecules and depicts the similarity of those features. It can be deducted that in those cases these pharmacophoric features bear special meaning when it comes to substrate or non-substrate classification. However, one should not forget that due to conformational optimisation for ROCS parameter calculation the training set had to be reduced and therefore results from one study to the other are not completely comparable. Nevertheless some deductions may be drawn.

Another compound that came into focus was compound 30 which represents a non-substrate compound. It was also named in the first study by descriptors 3D Autocorrelation and VSA and in the second study by ColorScore, ComboScore and ShapeTanimoto with binary QSAR. For random forest it held some impact for descriptors ColorScore, ComboScore, ShapeTanimoto, ScaledColor and VSA.

Other crucial substrate compounds of reference set D were compounds 1,4, 12, 14 and 15 but those have only been selected by either both classification methods in the second study or only by binary QSAR classification during both studies. Interestingly especially for random forest often VSA and VolSurf descriptors appointed the same compounds. These descriptors

seem to place importance onto the same structural features of the reference molecules. This is especially important for VSA descriptors with the random forest method as this combination together with reference set D rendered the best model so far (Figure 4.17).



(a) Compound 1 - S (b) Compound 4 - S (c) Compound 12 - S

(d) Compound 14 - S (e) Compound 15 - S (f) Compound 30 - NS

Figure 4.17: Further possible core structures of reference set D. S - Substrates, NS - Non-substrates.

Again interpretation of these results remains very difficult due to the complex and diverse nature of the compounds in the dataset. Also it has to be acknowledged that with only forty descriptors the possibility of chance encounters of the same compounds in the ten most important structures for model building when regarding two classification methods and nine different descriptor sets cannot be completely disclaimed. Therefore those structures have to be interpreted with caution though may still render some clues for substrate and non-substrate features when used with a shape similarity and SIBAR approach.

4.5 SIBAR approach

As similarity based methods maybe are as old as QSAR itself another method for similarity based classification remains the k-nearest neighbour (kNN) approach. In order to perceive the performance of SIBAR put against the kNN approach the compounds of the test set were classified by majority vote of their training set neighbours based on five ROCS parameters including ColorScore, ComboScore, Overlap, ScaledColor and ShapeTanimoto. The neighbours taken into account for activity profiling reached from one to five neighbours of the training set neighbours. Once again the results were mixed but in general the performance of SIBAR descriptors was significantly higher than the accuracies reached by the k-nearest neighbour approach. The overall accuracies of kNN reached from 52% (ShapeTanimoto) to 65% (Overlap) though with very low accuracy on substrates which ranged from 6 to 43%[260] (Table 4.20).

The main difference between these two methods lies in the fact that for kNN direct shape similarity is used for classification whereas SIBAR uses the shape similarity profile based on a set of 40 to 50 reference compounds as input structures. These results emphasise the validity of the SIBAR approach and show that in case of a set of compounds with high structural diversity and complexity the shape similarity approach outperforms individual shape comparison. However, it has to be added that the SIBAR approach renders SIBAR descriptors which are further analysed and tuned by various machine learning techniques whereas k-nearest neighbour represents a classification method in itself. kNN regards the average properties of the neighbours and based thereon proposes a probability value for a compound to be either substrate or non-substrate.

Seen in general the performance of the SIBAR approach gave satisfactory results though truly significant differences between SIBAR derived descriptors and the descriptors used in pure form could not be shown. Nevertheless the identification of core compounds for model building may provide fur-

Testset kNN	46 Comp. Descriptors	k	A	A1	A0	Pr1	Pr0
refB	ComboScore	1	52,17	62,50	46,67	38,46	70,00
	ComboScore	2	47,83	37,50	53,33	30,00	61,54
	ComboScore	3	45,65	43,75	46,67	30,43	60,87
	ComboScore	4	47,83	37,50	53,33	30,00	61,54
	ComboScore	5	47,83	43,75	50,00	31,82	62,50
refD	ComboScore	1	56,52	68,75	50,00	42,31	75,00
	ComboScore	2	65,22	50,00	73,33	50,00	73,33
	ComboScore	3	69,57	68,75	70,00	55,00	80,77
	ComboScore	4	63,04	37,50	76,67	46,15	69,70
	ComboScore	5	63,04	62,50	63,33	47,62	76,00

Table 4.20: Table 4.20 shows the results for the k-nearest neighbour approach. The number of nearest neighbours (k) reaches from 1 to 5, refB – reference set B, refD – reference set D, A – overall accuracy, A1 – accuracy on substrates, A0 – accuracy on non-substrates, Pr1 – precision on substrates, Pr0 – precision on non-substrates.

ther insights into ABCB1 substrate recognition mechanism and important features.

In order to better compare SIBAR approaches to pure descriptor results the mean overall accuracy over all descriptor sets for one method and each reference set was used. That means that the mean overall accuracy for each of the methods used was calculated. Consequently SIBAR reference set A, SIBAR reference set B, SIBAR reference set C, SIBAR reference set D and the pure descriptors were compared with one another (Table 4.21).

In the first study using WEKA support vector machine and binary QSAR SIBAR could not assert itself significantly. Using the support vector machine and the radial basis function kernel no difference in performance could be observed regarding the mean overall accuracy over all three descriptors and the single reference sets used. Especially reference set D showed poor results mainly concerning prediction accuracy on substrates which was down to 44 percent. Regarding only the pure descriptors all three of them showed robust

results and particularly 2D descriptors achieved overall accuracies of 75% with also very good accuracy on substrate values of nearly 89%.

Method Binary QSAR	Refset		A	A1	A0	Pr1	Pr0
	refA	MEAN	68,75	42,59	84,44	66,31	70,87
		STDEV	10,83	11,56	13,47	24,16	6,57
	refB	MEAN	66,67	42,59	81,11	57,22	70,74
		STDEV	4,17	19,51	9,62	8,55	5,16
	refC	MEAN	66,67	53,70	74,44	60,87	72,59
		STDEV	10,83	13,98	20,09	18,16	6,36
	refD	MEAN	63,19	27,78	84,44	48,52	66,80
		STDEV	3,18	24,22	9,62	7,88	5,70
	pure	MEAN	62,50	42,59	74,44	50,07	68,61
		STDEV	3,61	13,98	9,62	6,22	3,24

Table 4.21: Accuracies averaged over every descriptor set used with focus on the reference set. Results of Study 1.

A similar picture is drawn when regarding results of support vector machine with the polynomial kernel. Here although the best models regarding all three descriptor sets could be achieved with reference set B the pure descriptors without SIBAR did not do significantly worse. Again 2D descriptors alone as well as with SIBAR could produce very good rates of accuracy on substrates of around 78%.

Things went differently when regarding binary QSAR results of the first study. Here clearly the best results could be drawn with SIBAR opposed to pure descriptors alone. This remains true for the three satellite reference sets in particular. Regarding the three satellite reference sets based on all descriptor sets together each of the satellite reference sets could achieve a mean overall accuracy of 67 to 69% whereas the mean overall accuracy over all three pure descriptor sets moved along 62%. When the same procedure was applied to the data of study 2 similar results could be observed. The best SIBAR result could be found for reference set D (66.52%) in this case and the worst accuracy averaged over all descriptor sets used was given by

reference set B (60.87%). The difference to the averaged pure descriptor sets was slight (60.14%) (Figure 4.18).

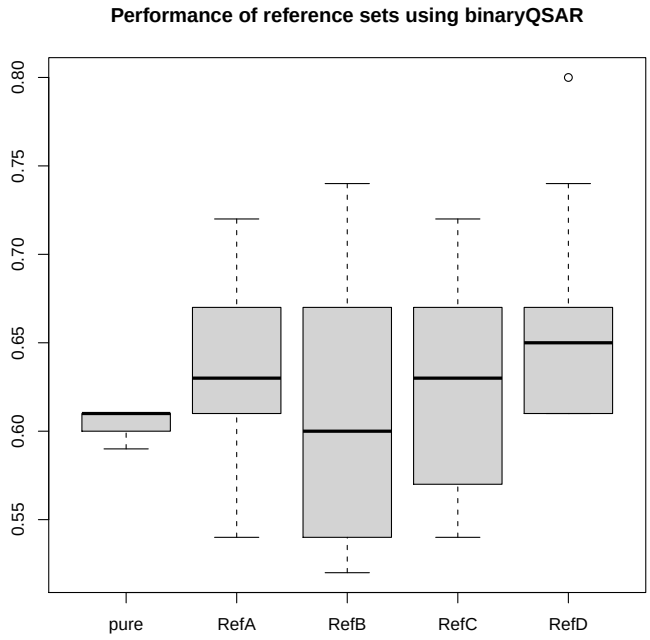


Figure 4.18: Accuracies with focus on reference sets and pure descriptors using binary QSAR. Results of study 2.

When looking again at the support vector machine of the second study similar results can be observed. The mean accuracies of all four SIBAR reference sets opposed to the pure descriptors reached higher overall prediction accuracies of 63.54% to 65.46% whereas the pure descriptors achieved a prediction accuracy of 61.59%. When regarding the performances of random forest models no clear preference can be named. In both random forest approaches the differences of the built models ranged from high to very low. That means that no clear conclusion can be drawn (Figure 4.19).

For the first random forest approach the average performance of the four reference sets (62.56%, 63.91%, 60.87%, 65.65%) lay below the average performance of the pure descriptors with 66.67%. This is highly interesting as the best model of the study could be built using random forest in combina-

tion with VSA descriptors and reference set D (82,61%). In spite of this high overall accuracy the other descriptor sets could not build similarly successful models and the average was lowered down to nearly 67%. The second random forest approach overset these results as in this case all four reference sets received better average accuracy from 60.63% to 65.22% whereas the pure descriptors achieved an overall accuracy of 59.42%. This shows again the fickleness of this comparative approach. When regarding the tuned random forest results change once again. Here the pure descriptors performed slightly better than SIBAR reference set C but worse than SIBAR reference sets A, B and D (Figure 4.20) .

In general with the exception of the fickle random forest the SIBAR derived models performed slightly better than the models developed with the pure descriptors. However it has to be granted that the number of SIBAR models was higher than the models of pure descriptor sets as the ROCS

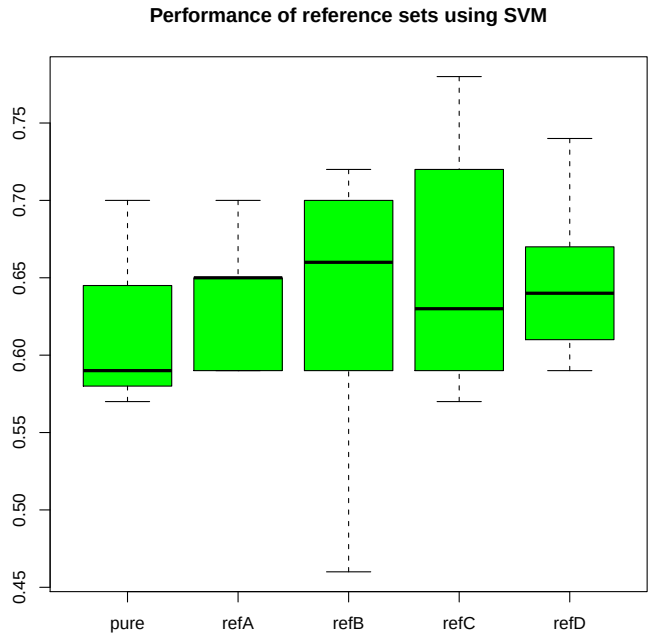


Figure 4.19: Accuracies with focus on reference sets and pure descriptors using support vector machine. Results of study 2.

parameters represent shape comparison values depending on two different shapes to compare. They have been taken into consideration for averaging anyway as they still represent a valid SIBAR like approach. Of course it has to be mentioned as well that the standard deviation averaging over nine to ten descriptor sets (including 10ADME descriptors) was higher than when averaging over only three pure descriptors including 10 ADME.

The probability of occurring outliers grows the more models are analysed. Bearing this in mind it has to be conceded that the performance of SIBAR and shape similarity did not bring the hoped for significant impact on model building for ABCB1. This may be due to the diversity of the dataset as the successful implementation of SIBAR has been shown in a number of studies of our group.[118,210,211] The dataset in this case as previously mentioned consists of highly complex natural products which increases the challenge that substrate prediction for so promiscuous a transporter as ABCB1 represents

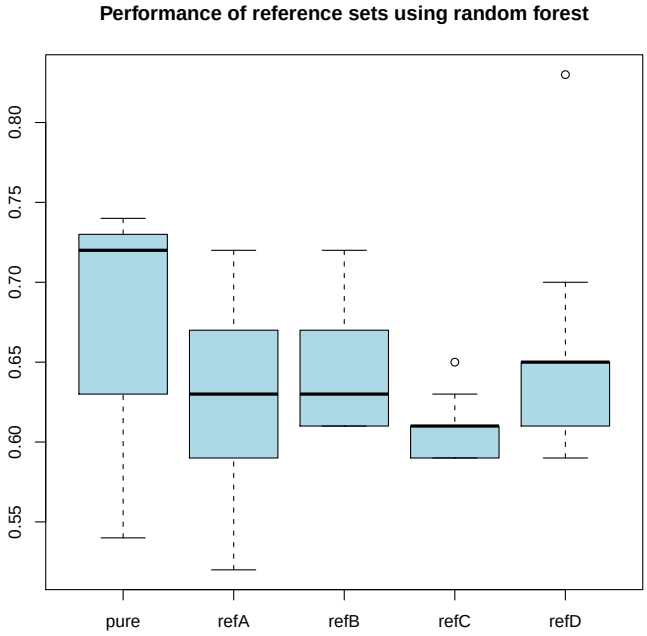


Figure 4.20: Accuracies with focus on reference sets and pure descriptors using random forest. Results of study 2.

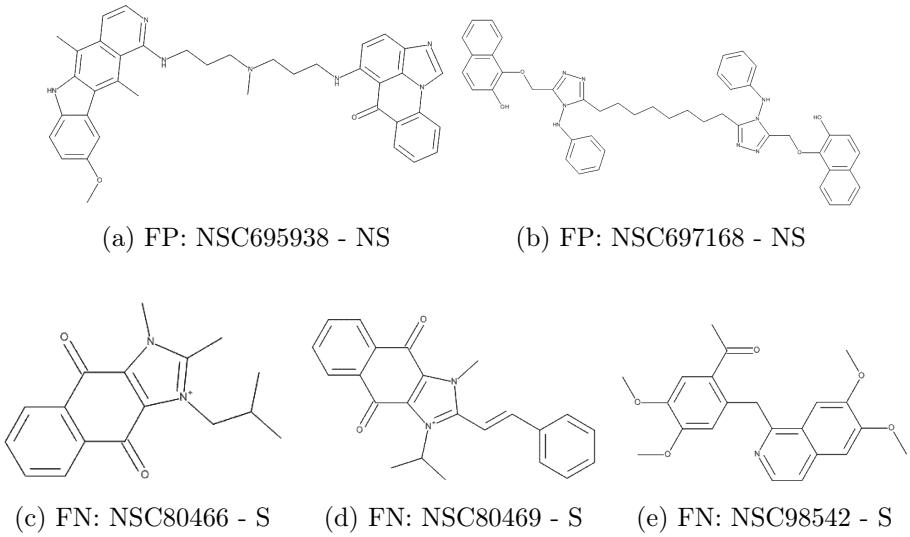
in itself.

4.6 Analysis of Misclassifications (taken from publication[260])

Compounds may be classified as False Positives (non-substrates predicted as substrate) and False Negatives (substrates predicted as non-substrates) for the following reasons: (i) only little representation of the respective chemical scaffold in the training set, (ii) representation of segments of the structure in the training set with different activity class assigned, (iii) a classification output close to the threshold between substrates (more than 0.5) and non-substrates (less than 0.5 predicted activity), (iv) a Pearson correlation coefficient close to the thresholds used for annotating substrates (-0,3). Within the next section selected misclassified compounds will be presented together with an analysis of possible reasons for misclassification.

4.6.1 False Positives

Compound NSC 695938 (Figure 4.21a) seems to be a heterodimer of two scaffolds active in tumour therapy and related to natural products. There are at least 14 molecules in the training set which are based on one of the two basic sub structures, and 13 of them are categorised as substrates. Thus, it seems likely that in this case the misclassification is mainly based on the high similarity of a certain sub-structure to a class of substrates rather than on a borderline Pearson coefficient (0.013). This seems also to be the case for compound NSC 697168 (Figure 4.21b) which is a homodimer and shows six partially similar structures in the training set. Four of them are classified as substrates and two are classified as non-substrates. The Pearson coefficient in this case is 0.0124089.



(a) FP: NSC695938 - NS (b) FP: NSC697168 - NS

(c) FN: NSC80466 - S (d) FN: NSC80469 - S (e) FN: NSC98542 - S

Figure 4.21: Compounds of the dataset most often misclassified. FP - False Positives: NSC 695938, NSC 697168; FN - False Negatives: NSC 80466, NSC 80469, NSC 98542. S – Substrate, NS – Non-substrates.

4.6.2 False Negatives

Concerning the non-substrates there are three molecules most often classified as non-substrates with all three classification methods.

Compound NSC 80466 (Figure 4.21c) and also Compound NSC 80469 (Figure 4.21d) are two relatively similar compounds with slight differences in their respective side chains. There is one molecule in the training set with a similar scaffold, which has an phenyl ring attached to the quaternary nitrogen atom and is correctly classified as substrate. As the Pearson correlation coefficient for both false negatives is far beyond the threshold (-0.467 for NSC 80466 and -0.638 for NSC 80469), we currently cannot provide a reasonable explanation for the misclassification of these two compounds. Another compound very often erroneously annotated as non-substrate is NSC 98542 (Figure 4.21e). The Pearson correlation coefficient in this case is -0.354, which is quite close to the threshold value and might be within the "grey zone".

5

Conclusion

First of all it has to be stated that the transporter ABCB1 remains a highly challenging task regarding model building and substrate prediction. This transporter is known and notorious for its high promiscuity and unpredictable substrate binding. In order to gain further insights similarity based methods have been applied throughout this study. In summary it can be said that models based on SIBAR descriptors or shape similarities produced acceptable prediction accuracies and were comparable to models built on conventionally derived descriptors.

Regarding classification methods several methods including binary QSAR, support vector machine and random forest have been compared. Overall random forest has performed best followed by the binary QSAR and support vector machine. Linear and quadratic discriminant analysis could not compare in performance and stability. This may be due to the immense structural diversity and complexity present in the natural product compounds of the dataset. The known promiscuity of the transporter adds to these difficulties making more complex methods such as nonlinear classification and machine learning better suited than more conventional classical methods such as linear discriminant and quadratic discriminant analysis.

Regarding the descriptor performance of 2D versus 3D or shape based descriptors the performance of 2D descriptors compared incredibly well with 3D derived descriptors. Especially when compared to 3D Autocorrelation

descriptors VSA and also the set of 2D descriptors showed overall better prediction accuracies. When matching the performance of VSA descriptors to 3D VolSurf and shape-based descriptors occasional outliers could be observed but the best overall model with 83% accuracy was based on SIBAR reference set D, VSA descriptors and the random forest classification method. This was put down to the fact that VSA descriptors nominally are 2D descriptors but encompass 3D information over the van der Waal surface area volume information. The van der Waal surface area is calculated in combination with molar refractivity, partial electronic charges and lipophilicity and therefore renders a realistic picture of the molecule's 3D information.

When inspecting the performance of the reference sets conflictive outcomes accompany that endeavour. In the first study the reference sets based on the satellite ChemGPS approach proposed by Oprea and colleagues[212] could prevail whereas in the second study the tailored reference set D accomplished higher accuracies. Similar results regarding reference set D have been published by Zdrazil and colleagues[118] where also the tailored reference set rendered best performances. Nevertheless also the general reference set based on satellite structures could keep up.

In this study certain core structures important for SIBAR model building could be identified and may provide further insight into ABCB1 substrate recognition. Therefore the selection of the reference set depends on the goal that should be accomplished. In order to get a quick look and an overview over one's data with similarity based methods a globally applicable reference set like reference set B is recommended. For further investigation and more intense interaction with one's data a tailored reference set derived out of the most diverse members of the examined dataset is suggested.

The aim of this study was to investigate the performance of shape based and similarity based methods. In summary the classical SIBAR methods showed good results and acceptable prediction accuracies. The same is true for the shape similarity approach implemented with ROCS. A few highly able parameters could be detected like the ColorScore or the Tversky index that can be successfully used for model development. The ColorScore combines molecular features with similarities between two compounds and thereby

seems to be an ideal marker for molecules. Though shape similarity as well as SIBAR descriptors performed satisfactorily no significant better performance compared to the descriptors used in pure form could be proven. Overall some of the achieved models of the study could compare well with other studies published in this regard though did not reach the high accuracy of for example Huang and colleagues.[85]

Bibliography

[1] Gerhard F. Ecker. In silico screening of promiscuous targets and antitargets. *Chemistry Today*, 22(3), 2005.

[2] Alfred H Schinkel and Johan W Jonker. Mammalian drug efflux transporters of the ATP binding cassette (ABC) family: an overview. *Advanced Drug Delivery Reviews*, 55(1):3–29, January 2003.

[3] Gloria Lee and Reina Bendayan. Functional expression and localization of p-glycoprotein in the central nervous system: Relevance to the pathogenesis and treatment of neurological disorders. *Pharmaceutical Research*, 21(8):1313–1330–1330, August 2004.

[4] Gergely Szakács, András Vàradi, Csilla Özvegy-Laczka, and Balázs Sarkadi. The role of ABC transporters in drug absorption, distribution, metabolism, excretion and toxicity (ADME-Tox). *Drug Discovery Today*, 13(9-10):379–393, May 2008.

[5] David S. Miller. Regulation of p-glycoprotein and other ABC drug transporters at the blood-brain barrier. *Trends in Pharmacological Sciences*, 31(6):246–254, June 2010.

[6] Martin F. Fromm. Importance of p-glycoprotein at blood-tissue barriers. *Trends in Pharmacological Sciences*, 25(8):423–429, August 2004.

[7] Alfred H. Schinkel. The physiological function of drug-transporting p-glycoproteins. *Seminars in Cancer Biology*, 8(3):161–170, June 1997.

[8] Wolfgang Löscher and Heidrun Potschka. Blood-brain barrier active efflux transporters: ATP-Binding cassette gene family. *The Journal of the American Society for Experimental NeuroTherapeutics*, 2(1):86–98, January 2005.

[9] Martin F. Fromm, Richard B. Kim, C. Michael Stein, Grant R. Wilkinson, and Dan M. Roden. Inhibition of p-glycoprotein-mediated drug transport : A unifying mechanism to explain the interaction between digoxin and quinidine. *Circulation*, 99(4):552 –557, February 1999.

[10] Christiane Pauli-Magnus and Peter J. Meier. Hepatobiliary transporters and drug-induced cholestasis. *Hepatology*, 44(4):778–787, 2006.

[11] Miriam Huls, Franz G. M. Russel, and Rosalind Masereeuw. The role of ATP binding cassette transporters in tissue defense and organ regeneration. *The Journal of Pharmacology and Experimental Therapeutics*, 328(1):3–9, January 2009.

[12] Heidrun Potschka. Targeting regulation of ABC efflux transporters in brain diseases: A novel therapeutic approach. *Pharmacology & Therapeutics*, 125(1):118–127, January 2010.

[13] Jens Pahnke, Olaf Wolkenhauer, Olaf, Markus Krohn, and Lary C. Walker. Clinico-pathologic function of cerebral ABC transporters - implications for the pathogenesis of alzheimer's disease. *Current Alzheimer Research*, 5(4):396–405, 2008.

[14] Enoche F Oga, Shuichi Sekine, Yoshihisa Shitara, and Toshiharu Horie. P-glycoprotein mediated efflux in caco-2 cell monolayers: The influence of herbals on digoxin transport. *Journal of ethnopharmacology*, 144(3):612–617, December 2012. PMID: 23064285.

[15] Membrane transporters in drug development. *Nat Rev Drug Discov*, 9(3):215–236, March 2010.

[16] Olga Wesołowska. Interaction of phenothiazines, stilbenes and flavonoids with multidrug resistance-associated transporters, p-glycoprotein and MRP1. *Acta biochimica Polonica*, 58(4):433–448, 2011. PMID: 22187677.

[17] Frances J Sharom. The p-glycoprotein multidrug transporter. *Essays in biochemistry*, 50(1):161–178, September 2011. PMID: 21967057.

[18] Thomas J. Raub. P-glycoprotein recognition of substrates and circumvention through rational drug design. *Molecular Pharmaceutics*, 3(1):3–25, February 2006.

[19] M. Bebawy and M. Chetty. Gender differences in p-glycoprotein expression and function: Effects on drug disposition and outcome. *Current Drug Metabolism*, 10(4):322–328, May 2009.

[20] Mark F. Rosenberg, Giles Velarde, Robert C. Ford, Catherine Martin, Georgina Berridge, Ian D. Kerr, Richard Callaghan, Andreas Schmidlin, Carol Wooding, Kenneth J. Linton, and Christopher F. Higgins. Repacking of the transmembrane domains of p-glycoprotein during the transport ATPase cycle. *EMBO J*, 20(20):5615–5625, October 2001.

[21] Mark F. Rosenberg, Alhaji Bukar Kamis, Richard Callaghan, Christopher F. Higgins, and Robert C. Ford. Three-dimensional structures of the mammalian multidrug resistance p-glycoprotein demonstrate major conformational changes in the transmembrane domains upon nucleotide binding. *Journal of Biological Chemistry*, 278(10):8294 –8299, March 2003.

[22] Roger J. P. Dawson, Kaspar Hollenstein, and Kaspar P. Locher. Uptake or extrusion: crystal structures of full ABC transporters suggest a common mechanism. *Molecular Microbiology*, 65(2):250–257, 2007.

[23] Douglas C. Rees, Eric Johnson, and Oded Lewinson. ABC transporters: the power to change. *Nat Rev Mol Cell Biol*, 10(3):218–227, March 2009.

[24] Freya Klepsch and Gerhard F. Ecker. Impact of the recent mouse p-glycoprotein structure for structure-based ligand design. *Molecular Informatics*, 29(4):276–286, 2010.

[25] Geoffrey Chang, Christopher B Roth, Christopher L Reyes, Owen Pornillos, Yen-Ju Chen, and Andy P Chen. Retraction. *Science (New York, N.Y.)*, 314(5807):1875, December 2006. PMID: 17185584.

[26] Roger J. P. Dawson and Kaspar P. Locher. Structure of a bacterial multidrug ABC transporter. *Nature*, 443(7108):180–185, 2006.

[27] Tip W. Loo, M. Claire Bartlett, and David M. Clarke. Substrate-induced conformational changes in the transmembrane segments of human p-glycoprotein. *Journal of Biological Chemistry*, 278(16):13603 –13606, April 2003.

[28] Kaspar Hollenstein, Roger JP Dawson, and Kaspar P Locher. Structure and mechanism of ABC transporter proteins. *Current Opinion in Structural Biology*, 17(4):412–418, August 2007.

[29] Andrew Ward, Christopher L. Reyes, Jodie Yu, Christopher B. Roth, and Geoffrey Chang. Flexibility in the ABC transporter MsbA: alternating access with a twist. *Proceedings of the National Academy of Sciences*, 104(48):19005 –19010, November 2007.

[30] Tip W. Loo, M. Claire Bartlett, and David M. Clarke. Human p-glycoprotein is active when the two halves are clamped together in the closed conformation. *Biochemical and Biophysical Research Communications*, 395(3):436–440, May 2010.

[31] Brandy Verhalen and Stephan Wilkens. P-glycoprotein retains drug-stimulated ATPase activity upon covalent linkage of the two nucleotide binding domains at their c-terminal ends. *Journal of Biological Chemistry*, 286(12):10476 –10482, March 2011.

[32] Stephen G. Aller, Jodie Yu, Andrew Ward, Yue Weng, Srinivas Chittaboina, Rupeng Zhuo, Patina M. Harrell, Yenphuong T. Trinh, Qinghai Zhang, Ina L. Urbatsch, and Geoffrey Chang. Structure of p-glycoprotein reveals a molecular basis for poly-specific drug binding. *Science*, 323(5922):1718 –1722, March 2009.

[33] Mi Sun Jin, Michael L Oldham, Qiuju Zhang, and Jue Chen. Crystal structure of the multidrug transporter p-glycoprotein from caenorhabditis elegans. *Nature*, 490(7421):566–569, October 2012. PMID: 23000902.

[34] Mark F. Rosenberg, Richard Callaghan, Robert C. Ford, and Christopher F. Higgins. Structure of the multidrug resistance p-glycoprotein to 2.5 nm resolution determined by electron microscopy and image analysis. *Journal of Biological Chemistry*, 272(16):10685 –10694, April 1997.

[35] Frances J. Sharom. Shedding light on drug transport: structure and function of the p-glycoprotein multidrug transporter (ABCB1)This paper is one of a selection of papers published in this special issue, entitled CSBMCB — membrane proteins in health and disease. *Biochem. Cell Biol.*, 84(6):979–992, 2006.

[36] Tip W. Loo, M. Claire Bartlett, and David M. Clarke. Simultaneous binding of two different drugs in the binding pocket of the human multidrug resistance p-glycoprotein. *Journal of Biological Chemistry*, 278(41):39706 –39710, October 2003.

[37] Christopher F Higgins and Kenneth J Linton. The ATP switch model for ABC transporters. *Nat Struct Mol Biol*, 11(10):918–926, October 2004.

[38] Daniel A.P. Gutmann, Andrew Ward, Ina L. Urbatsch, Geoffrey Chang, and Hendrik W. van Veen. Understanding polyspecificity of multidrug ABC transporters: closing in on the gaps in ABCB1. *Trends in Biochemical Sciences*, 35(1):36–42, January 2010.

[39] Kaspar Hollenstein, Dominik C. Frei, and Kaspar P. Locher. Structure of an ABC transporter in complex with its binding protein. *Nature*, 446(7132):213–216, March 2007.

[40] Markus A. Seeger and Hendrik W. van Veen. Molecular basis of multidrug transport by ABC transporters. *Biochimica et Biophysica Acta (BBA) - Proteins & Proteomics*, 1794(5):725–737, May 2009.

[41] Robert Ernst, Petra Kueppers, Jan Stindt, Karl Kuchler, and Lutz Schmitt. Multidrug efflux pumps: Substrate selection in ATP-binding cassette multidrug efflux pumps – first come, first served? *FEBS Journal*, 277(3):540–549, 2010.

[42] John G Wise. Catalytic transitions in the human MDR1 p-glycoprotein drug binding sites. *Biochemistry*, 51(25):5125–5141, June 2012. PMID: 22647192.

[43] Ming Liu, Tingjun Hou, Zhiwei Feng, and Youyong Li. The flexibility of p-glycoprotein for its poly-specific drug binding from molecular dynamics simulations. *Journal of biomolecular structure $0 dynamics*, August 2012. PMID: 22888853.

[44] Ian D. Kerr, Peter M. Jones, and Anthony M. George. Multidrug efflux pumps: The structures of prokaryotic ATP-binding cassette transporter efflux pumps and implications for our understanding of eukaryotic p-glycoproteins and homologues. *FEBS Journal*, 277(3):550–563, 2010.

[45] Michael A Demel, R Schwaha, O Krämer, P Ettmayer, Eric EJ Haaksma, and Gerhard F Ecker. In silico prediction of substrate properties for ABC-multidrug transporters. *Expert Opin. Drug Metab. Toxicol.*, 4(9):1167–1180, August 2008.

[46] Terry R. Stouch and Olafur Gudmundsson. Progress in understanding the structure-activity relationships of p-glycoprotein. *Advanced Drug Delivery Reviews*, 54(3):315–328, March 2002.

[47] Hongyu Zhou, Shuhong Wu, Shumei Zhai, Aifeng Liu, Ying Sun, Rongshi Li, Ying Zhang, Sean Ekins, Peter W. Swaan, Bingliang Fang, Bin Zhang, and Bing Yan. Design, synthesis, cytoselective toxicity, Structure–Activity relationships, and pharmacophore of thiazolidinone derivatives targeting drug-resistant lung cancer cells. *Journal of Medicinal Chemistry*, 51(5):1242–1251, March 2008.

[48] Bo Feng, Jessica B. Mills, Ralph E. Davidson, Rouchelle J. Mireles, John S. Janiszewski, Matthew D. Troutman, and Sonia M. de Morais. In vitro p-glycoprotein assays to predict the in vivo interactions of p-glycoprotein with drugs in the central nervous system. *Drug Metabolism and Disposition*, 36(2):268 –275, February 2008.

[49] Joseph W. Polli, Stephen A. Wring, Joan E. Humphreys, Liyue Huang, Jonathon B. Morgan, Lindsey O. Webster, and Cosette S. Serabjit-Singh. Rational use of in vitro p-glycoprotein assays in drug discovery. *Journal of Pharmacology and Experimental Therapeutics*, 299(2):620 –628, November 2001.

[50] Jerome H. Hochman, Masayo Yamazaki, Tomoyuki Ohe, and Jiunn H. Lin. Evaluation of drug interactions with p-glycoprotein in drug discovery: In vitro assessment of the potential for drug-drug interactions with p-glycoprotein. *Current Drug Metabolism*, 3(3):257–273, 2002.

[51] Pierluigi Nervi, Xiaochun Li-Blatter, Päivi Äanismaa, and Anna Seelig. P-glycoprotein substrate transport assessed by comparing cellular and vesicular ATPase activity. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1798(3):515–525, March 2010.

[52] Cheng Chang and Peter W. Swaan. Computational approaches to modeling drug transporters. *European Journal of Pharmaceutical Sciences*, 27(5):411–424, April 2006.

[53] Christopher A Lipinski, Franco Lombardo, Beryl W Dominy, and Paul J Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Special issue dedicated to Dr. Eric Tomlinson, Advanced Drug Delivery Reviews, A Selection of the Most Highly Cited Articles, 1991-1998*, 46(1–3):3–26, March 2001.

[54] M Paul Gleeson. Generation of a set of simple, interpretable ADMET rules of thumb. *Journal of medicinal chemistry*, 51(4):817–834, February 2008. PMID: 18232648.

[55] Stephen A Hitchcock. Structural modifications that alter the p-glycoprotein efflux properties of compounds. *Journal of medicinal chemistry*, 55(11):4877–4895, June 2012. PMID: 22506484.

[56] H L Pearce, A R Safa, N J Bach, M A Winter, M C Cirtain, and W T Beck. Essential features of the p-glycoprotein pharmacophore as defined by a series of reserpine analogs that modulate multidrug resistance. *Proceedings of the National Academy of Sciences*, 86(13):5128 –5132, July 1989.

[57] G. Ecker, M. Huber, D. Schmid, and P. Chiba. The importance of a nitrogen atom in modulators of multidrug resistance. *Molecular Pharmacology*, 56(4):791 –796, October 1999.

[58] Anna Seelig. A general pattern for substrate recognition by p-glycoprotein. *European Journal of Biochemistry*, 251(1-2):252–261, 1998.

[59] Anna Seelig and Ewa Landwojtowicz. Structure-activity relationship of p-glycoprotein substrates and modifiers. *European Journal of Pharmaceutical Sciences*, 12(1):31–40, November 2000.

[60] Gerhard Wolber, Thomas Seidel, Fabian Bendix, and Thierry Langer. Molecule-pharmacophore superpositioning and pattern matching in computational drug design. *Drug Discovery Today*, 13(1-2):23–29, January 2008.

[61] Cheng Chang, Sean Ekins, Praveen Bahadduri, and Peter W. Swaan. Pharmacophore-based discovery of ligands for drug transporters. *Advanced Drug Delivery Reviews*, 58(12-13):1431–1450, November 2006.

[62] Julie E. Penzotti, Michelle L. Lamb, Erik Evensen, and Peter D. J. Grootenhuis. A computational ensemble pharmacophore model for identifying substrates of p-glycoprotein. *Journal of Medicinal Chemistry*, 45(9):1737–1740, April 2002.

[63] Sean Ekins, Richard B. Kim, Brenda F. Leake, Anne H. Dantzig, Erin G. Schuetz, Lu-Bin Lan, Kazuto Yasuda, Robert L. Shepard, Mark A. Winter, John D. Schuetz, James H. Wikel, and Steven A. Wrighton. Three-dimensional quantitative structure-activity relationships of inhibitors of p-glycoprotein. *Molecular Pharmacology*, 61(5):964 –973, May 2002.

[64] Sean Ekins, Richard B. Kim, Brenda F. Leake, Anne H. Dantzig, Erin G. Schuetz, Lu-Bin Lan, Kazuto Yasuda, Robert L. Shepard, Mark a Winter, John D. Schuetz, James H. Wikel, and Steven A. Wrighton. Application of three-dimensional quantitative structure-activity relationships of p-glycoprotein inhibitors and substrates. *Molecular Pharmacology*, 61(5):974 –981, May 2002.

[65] Sean Ekins, Gianpaolo Bravi, James H. Wikel, and Steven A. Wrighton. Three-dimensional-quantitative structure activity relationship analysis of cytochrome p-450 3A4 substrates. *Journal of Pharmacology and Experimental Therapeutics*, 291(1):424 –433, October 1999.

[66] Sean Ekins, Gianpaolo Bravi, Shelly Binkley, Jennifer S. Gillespie, Barbara J. Ring, James H. Wikel, and Steven A Wrighton. Three- and four-dimensional quantitative structure activity relationship analyses of cytochrome p-450 3A4 inhibitors. *Journal of Pharmacology and Experimental Therapeutics*, 290(1):429 –438, July 1999.

[67] Cheng Chang, Praveen M. Bahadduri, James E. Polli, Peter W. Swaan, and Sean Ekins. Rapid identification of p-glycoprotein substrates and inhibitors. *Drug Metabolism and Disposition*, 34(12):1976 –1984, December 2006.

[68] Ilza K. Pajeva and Michael Wiese. Pharmacophore model of drugs involved in p-glycoprotein multidrug resistance: Explanation of structural variety (hypothesis). *Journal of Medicinal Chemistry*, 45(26):5671–5686, December 2002.

[69] Alexia Garrigues, Nicolas Loiseau, Marcel Delaforge, Jacques Ferté, Manuel Garrigos, François André, and Stéphane Orlowski. Characterization of two pharmacophores on the multidrug transporter p-glycoprotein. *Molecular Pharmacology*, 62(6):1288 –1298, December 2002.

[70] Giovanni Cianchetta, Robert W. Singleton, Meng Zhang, Marianne Wildgoose, Dennis Giesing, Arnaldo Fravolini, Gabriele Cruciani, and Roy J. Vaz. A pharmacophore hypothesis for p-glycoprotein substrate recognition using GRIND-Based 3D-QSAR. *Journal of Medicinal Chemistry*, 48(8):2927–2935, April 2005.

[71] Patrizia Crivori, Benedetta Reinach, Daniele Pezzetta, and Italo Poggesi. Computational models for identifying potential p-glycoprotein substrates and inhibitors. *Molecular Pharmaceutics*, 3(1):33–44, February 2006.

[72] Wu-Xiong Li, Leping Li, John Eksterowicz, Xuefeng Bruce Ling, and Mario Cardozo. Significance analysis and multiple pharmacophore models for differentiating p-glycoprotein substrates. *Journal of Chemical Information and Modeling*, 47(6):2429–2438, November 2007.

[73] Elena Dolghih, Clifford Bryant, Adam R Renslo, and Matthew P Jacobson. Predicting binding to p-glycoprotein by flexible receptor docking. *PLoS computational biology*, 7(6):e1002083, June 2011. PMID: 21731480.

[74] Remigijus Didziapetris, Pranas Japertas, Alex Avdeef, and Alanas Petrauskas. Classification analysis of p-glycoprotein substrate specificity. *Journal of Drug Targeting*, 11(7):391–406, 2003.

[75] Pharma-Algorithms. ADME boxes. http://pharma-algorithms.com/adme_boxes.htm, 2011.

[76] Remigijus Didziapetris, Pranas Japertas, Laurynas Riauba, and Alanas Petrauskas. Classification SAR (c-SAR) in prediction of p-glycoprotein substrate specificity. *Poster EuroQSAR*, 2002.

[77] Marc Adenot and Roger Lahana. Blood-brain barrier permeation models: Discriminating between potential CNS and non-CNS drugs including p-glycoprotein substrates. *Journal of Chemical Information and Computer Sciences*, 44(1):239–248, January 2004.

[78] Vijay K. Gombar, Joseph W. Polli, Joan E. Humphreys, Stephen A. Wring, and Cosette S. Serabjit-Singh. Predicting p-glycoprotein substrates by a quantitative structure–activity relationship model. *Journal of Pharmaceutical Sciences*, 93(4):957–968, 2004.

[79] Y. Xue, C. W. Yap, L. Z. Sun, Z. W. Cao, J. F. Wang, and Y. Z. Chen. Prediction of p-glycoprotein substrates by a support vector machine approach. *Journal of Chemical Information and Computer Sciences*, 44(4):1497–1505, July 2004.

[80] Y. Xue, Z. R. Li, C. W. Yap, L. Z. Sun, X. Chen, and Y. Z. Chen. Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents. *Journal of Chemical Information and Computer Sciences*, 44(5):1630–1638, 2004.

[81] Yong-Hua Wang, Yan Li, Sheng-Li Yang, and Ling Yang. Classification of substrates and inhibitors of p-glycoprotein using unsupervised machine learning approach. *Journal of Chemical Information and Modeling*, 45(3):750–757, May 2005.

[82] Tomasz Arodź, David A. Yuen, and Arkadiusz Z. Dudek. Ensemble of linear models for predicting drug properties. *Journal of Chemical Information and Modeling*, 46(1):416–423, January 2006.

[83] Miguel Angel Cabrera, Isabel González, Carlos Fernández, Carmen Navarro, and Marival Bermejo. A topological substructural approach for the prediction of p-glycoprotein substrates. *Journal of Pharmaceutical Sciences*, 95(3):589–606, 2006.

[84] Patricia de Cerqueira Lima, Alexander Golbraikh, Scott Oloff, Yunde Xiao, and Alexander Tropsha. Combinatorial QSAR modeling of p-glycoprotein substrates. *Journal of Chemical Information and Modeling*, 46(3):1245–1254, May 2006.

[85] Jianping Huang, Guangli Ma, Ishtiaq Muhammad, and Yiyu Cheng. Identifying p-glycoprotein substrates using a support vector machine optimized by a particle swarm. *Journal of Chemical Information and Modeling*, 47(4):1638–1647, July 2007.

[86] Litai Zhang, Praveen V. Balimane, Stephen R. Johnson, and Saeho Chong. Development of an in silico model for predicting efflux substrates in caco-2 cells. *International Journal of Pharmaceutics*, 343(1-2):98–105, October 2007.

[87] Sheng-Yong Yang, Qi Huang, Lin-Li Li, Chang-Ying Ma, Hui Zhang, Ru Bai, Qi-Zhi Teng, Ming-Li Xiang, and Yu-Quan Wei. An integrated scheme for feature selection and parameter setting in the support vector machine modeling and its application to the prediction of pharmacokinetic properties of drugs. *Artificial Intelligence in Medicine*, 46(2):155–163, June 2009.

[88] Zhi Wang, Yuanying Chen, Hu Liang, Andreas Bender, Robert C. Glen, and Aixia Yan. P-glycoprotein substrate models using support vector machines based on a comprehensive data set. *Journal of Chemical Information and Modeling*, 51(6):1447–1456, May 2011.

[89] Fabio Broccatelli. QSAR models for p-glycoprotein transport based on a highly consistent data set. *Journal of chemical information and modeling*, 52(9):2462–2470, September 2012. PMID: 22946765.

[90] Vasanthanathan Poongavanam, Norbert Haider, and Gerhard F Ecker. Fingerprint-based in silico models for the prediction of p-glycoprotein substrates and inhibitors. *Bioorganic & medicinal chemistry*, 20(18):5388–5395, September 2012. PMID: 22595422.

[91] Fabio Broccatelli, Emanuele Carosati, Annalisa Neri, Maria Frosini, Laura Goracci, Tudor I Oprea, and Gabriele Cruciani. A novel approach for predicting p-glycoprotein (ABCB1) inhibition using molecular interaction fields. *Journal of medicinal chemistry*, 54(6):1740–1751, March 2011. PMID: 21341745.

[92] Chunbo Zhang, Patrick Kwan, Zhong Zuo, and Larry Baum. The transport of antiepileptic drugs by p-glycoprotein. *Advanced drug delivery reviews*, 64(10):930–942, July 2012. PMID: 22197850.

[93] Hugo Kubinyi. Chemical similarity and biological activities. *Journal of the Brazilian Chemical Society*, 13:717–726, 2002.

[94] Gisbert Schneider, Petra Schneider, and Steffen Renner. Scaffold-hopping: How far can you jump? *QSAR & Combinatorial Science*, 25(12):1162–1171, 2006.

[95] Anthony Nicholls, Georgia B. McGaughey, Robert P. Sheridan, Andrew C. Good, Gregory Warren, Magali Mathieu, Steven W. Muchmore, Scott P. Brown, J. Andrew Grant, James A. Haigh, Neysa Nevins, Ajay N. Jain, and Brian Kelley. Molecular shape and medicinal chemistry: A perspective. *Journal of Medicinal Chemistry*, 53(10):3862–3886, February 2010.

[96] Adrian M Schreyer and Tom Blundell. USRCAT: real-time ultrafast shape recognition with pharmacophoric constraints. *Journal of Cheminformatics*, 4(1):27, 2012.

[97] J Mestres, D C Rohrer, and G M Maggiora. A molecular field-based similarity approach to pharmacophoric pattern recognition. *Journal of molecular graphics & modelling*, 15(2):114–121, 103–106, April 1997. PMID: 9385558.

[98] Miquel de Càceres, Jordi Villà, Juan J. Lozano, and Ferran Sanz. MIPSIM: similarity analysis of molecular interaction potentials. *Bioinformatics*, 16(6):568–569, June 2000.

[99] Rita Schwaha and Gerhard F Ecker. The similarity principle – new trends and applications in ligand-based drug discovery and ADMET profiling. *Scientia Pharmaceutica*, 76(1):5–18, 2008.

[100] ROCS | OpenEye scientific software. http://www.eyesopen.com/rocs.

[101] Schrödinger - product suites - list of all products - phase. http://www.schrodinger.com/productpage/14/13/.

[102] Steven L. Dixon, Alexander M. Smondyrev, and Shashidhar N. Rao. PHASE: a novel approach to pharmacophore modeling and 3D database searching. *Chemical Biology & Drug Design*, 67(5):370–372, 2006.

[103] David A. Evans, Thompson N. Doman, David A. Thorner, and Michael J. Bodkin. 3D QSAR methods: Phase and catalyst compared. *Journal of Chemical Information and Modeling*, 47(3):1248–1257, May 2007.

[104] Xiaofeng Liu, Hualiang Jiang, and Honglin Li. SHAFTS: a hybrid approach for 3D molecular similarity calculation. 1. method and assessment of virtual screening. *Journal of Chemical Information and Modeling*, 51(9):2372–2385, August 2011.

[105] Adel Hamza, Ning-Ning Wei, Ce Hao, Zhilong Xiu, and Chang-Guo Zhan. A novel and efficient ligand-based virtual screening approach using the HWZ scoring function and an enhanced shape-density model. *Journal of Biomolecular Structure and Dynamics*, pages 1–15, November 2012.

[106] Ewgenij Proschak, Matthias Rupp, Swetlana Derksen, and Gisbert Schneider. Shapelets: possibilities and limitations of shape-based virtual screening. *Journal of computational chemistry*, 29(1):108–114, January 2008. PMID: 17516427.

[107] James H Nettles, Jeremy L Jenkins, Chris Williams, Alex M Clark, Andreas Bender, Zhan Deng, John W Davies, and Meir Glick. Flexible 3D pharmacophores as descriptors of dynamic biological space. *Journal of molecular graphics & modelling*, 26(3):622–633, October 2007. PMID: 17395510.

[108] Fabien Fontaine, Evan Bolton, Yulia Borodina, and Stephen H Bryant. Fast 3D shape screening of large chemical databases through alignment-recycling. *Chemistry Central journal*, 1:12, 2007. PMID: 17880744.

[109] Elisabet Gregori-Puigjané and Jordi Mestres. SHED: shannon entropy descriptors from topological feature distributions. *Journal of chemical information and modeling*, 46(4):1615–1622, August 2006. PMID: 16859293.

[110] David Vidal, Michael Thormann, and Miquel Pons. LINGO, an efficient holographic text based method to calculate biophysical properties and intermolecular similarities. *Journal of chemical information and modeling*, 45(2):386–393, April 2005. PMID: 15807504.

[111] Chaoqian Cai, Jiayu Gong, Xiaofeng Liu, Hualiang Jiang, Daqi Gao, and Honglin Li. A novel, customizable and optimizable parameter method using spherical harmonics for molecular shape similarity comparisons. *Journal of Molecular Modeling*, 18(4):1597–1610, April 2012.

[112] Sunghwan Kim, Evan Bolton, and Stephen Bryant. PubChem3D: shape compatibility filtering using molecular shape quadrupoles. *Journal of Cheminformatics*, 3(1):25, 2011.

[113] Darryl Reid, Bashir S Sadjad, Zsolt Zsoldos, and Aniko Simon. LASSO-ligand activity by surface similarity order: a new tool for ligand based virtual screening. *Journal of computer-aided molecular design*, 22(6-7):479–487, July 2008. PMID: 18204980.

[114] Simona Distinto, Francesca Esposito, Johannes Kirchmair, M. Cristina Cardia, Marco Gaspari, Elias Maccioni, Stefano Alcaro, Patrick Markt, Gerhard Wolber, Luca Zinzula, and Enzo Tramontano. Identification of HIV-1 reverse transcriptase dual inhibitors by a combined shape-, 2D-fingerprint- and

pharmacophore-based virtual screening approach. *European Journal of Medicinal Chemistry*, 50(0):216–229, April 2012.

[115] Georgia B McGaughey, Robert P Sheridan, Christopher I Bayly, J Chris Culberson, Constantine Kreatsoulas, Stacey Lindsley, Vladimir Maiorov, Jean-Francois Truchon, and Wendy D Cornell. Comparison of topological, shape, and docking methods in virtual screening. *Journal of chemical information and modeling*, 47(4):1504–1519, August 2007. PMID: 17591764.

[116] Denis Fourches, Eugene Muratov, and Alexander Tropsha. Trust, but verify: On the importance of chemical structure curation in cheminformatics and QSAR modeling research. *Journal of Chemical Information and Modeling*, 50(7):1189–1204, July 2010.

[117] Christian Klein, Dominik Kaiser, Stephan Kopp, Peter Chiba, and Gerhard F Ecker. Similarity based SAR (SIBAR) as tool for early ADME profiling. *Journal of computer-aided molecular design*, 16(11):785–793, November 2002. PMID: 12825790.

[118] Barbara Zdrazil, Dominik Kaiser, Stephan Kopp, Peter Chiba, and Gerhard F. Ecker. Similarity-based descriptors (SIBAR) as tool for QSAR studies on p-glycoprotein inhibitors: Influence of the reference set. *QSAR & Combinatorial Science*, 26(5):669–678, 2007.

[119] Gergely Szakács, Jean-Philippe Annereau, Samir Lababidi, Uma Shankavaram, Angela Arciello, Kimberly J. Bussey, William Reinhold, Yanping Guo, Gary D. Kruh, Mark Reimers, John N. Weinstein, and Michael M. Gottesman. Predicting drug sensitivity and resistance. *Cancer cell*, 6(2):129–137, 2004.

[120] Uwe Scherf, Douglas T. Ross, Mark Waltham, Lawrence H. Smith, Jae K. Lee, Lorraine Tanabe, Kurt W. Kohn, William C. Reinhold, Timothy G. Myers, Darren T. Andrews, Dominic A. Scudiero, Michael B. Eisen, Edward A. Sausville, Yves Pommier, David Botstein, Patrick O. Brown, and John N. Weinstein. A gene expression database for the molecular pharmacology of cancer. *Nat Genet*, 24(3):236–244, March 2000.

[121] Alexander Tropsha. Best practices for QSAR model development, validation, and exploitation. *Molecular Informatics*, 29(6-7):476–488, 2010.

[122] Alexander Tropsha, Paola Gramatica, and Vijay K. Gombar. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR & Combinatorial Science*, 22(1):69–77, 2003.

[123] A Golbraikh and A Tropsha. Beware of q2! *Journal of molecular graphics & modelling*, 20(4):269–276, January 2002.

[124] Alexander Golbraikh, Min Shen, Zhiyan Xiao, Yun-De Xiao, Kuo-Hsiung Lee, and Alexander Tropsha. Rational selection of training and test sets for the development of validated QSAR models. *Journal of Computer-Aided Molecular Design*, 17(2-4):241–253, February 2003.

[125] Paola Gramatica. Principles of QSAR models validation: internal and external. *QSAR & Combinatorial Science*, 26(5):694–701, 2007.

[126] Paola Gramatica. A short history of QSAR evolution. http://www.qsarworld.com/Temp_Fileupload/Shorthistoryofqsar.pdf, 2011.

[127] David D. Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. In *Proceedings of the 10th European Conference on Machine Learning*, pages 4–15. Springer-Verlag, 1998.

[128] Harry Zhang. The optimality of naive bayes. In Valerie Barr and Zdravko Markov, editors, *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004)*. AAAI Press, 2004.

[129] Anthony E Klon, Jeffrey F Lowrie, and David J Diller. Improved naïve bayesian modeling of numerical data for absorption, distribution, metabolism and excretion (ADME) property prediction. *Journal of chemical information and modeling*, 46(5):1945–1956, October 2006. PMID: 16995725.

[130] David J. Hand and Keming Yu. Idiot's bayes: Not so stupid after all? *International Statistical Review / Revue Internationale de Statistique*, 69(3):385–398, December 2001. ArticleType: research-article / Full publication date: Dec., 2001 / Copyright © 2001 International Statistical Institute (ISI).

[131] Irina Rish. An empirical study of the naive bayes classifier. In *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, pages 41–46, 2001.

[132] M. Pazzani P. Domingos. Beyond independence: Conditions for the optimality of the simple baysian classifier. Technical report, University of California, Irvine, Irvine, CA 92717.

[133] Elias Gyftodimos and Peter A. Flach. Hierarchical bayesian networks: an approach to classification and learning for structured data. In *Department of Informatics, University of Szeged*, page 291–300. Springer, 2004.

[134] Zdzisław Pawlak. Rough sets, decision algorithms and bayes' theorem. *European Journal of Operational Research*, 136(1):181–189, January 2002.

[135] Peter Cheeseman, Matthew Self, Jim Kelly, Will Taylor, Don Freeman, and John Stutz. "Bayesian classification". In *Proceedings of the Seventh National Conference of Artificial Intelligence (AAAI-88)*, pages 607–611. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

[136] P Labute. Binary QSAR: a new method for the determination of quantitative structure activity relationships. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 444–455, 1999. PMID: 10380218.

[137] Chemical Computing Group. Molecular operating environment - MOE. http://www.chemcomp.com/.

[138] Hongmao Sun. A naive bayes classifier for prediction of multidrug resistance reversal activity on the basis of atom typing. *Journal of Medicinal Chemistry*, 48(12):4031–4039, May 2005.

[139] Jonathon Shiens. A tutorial on principal component analysis. Tutorial, Center for Neural Science, New York University, New York [etc.], 2009.

[140] Lindsay I Smith. A tutorial on principal components analysis. Technical report, 2002.

[141] Hongmao Sun. An accurate and interpretable bayesian classification model for prediction of hERG liability. *ChemMedChem*, 1(3):315–322, 2006.

[142] Francesca Demichelis, Paolo Magni, Paolo Piergiorgi, Mark A Rubin, and Riccardo Bellazzi. A hierarchical naïve bayes model for handling sample heterogeneity in classification problems: an application to tissue microarrays. *BMC bioinformatics*, 7:514, 2006. PMID: 17125514.

[143] Hua Gao, Chris Williams, Paul Labute, and Jürgen Bajorath. Binary quantitative Structure-Activity relationship (QSAR) analysis of estrogen receptor ligands. *Journal of Chemical Information and Computer Sciences*, 39(1):164–168, December 1998.

[144] Paul Watson. Naive bayes classification using 2D pharmacophore feature triplet vectors. *Journal of Chemical Information and Modeling*, 48(1):166–178, January 2008.

[145] Khac-Minh Thai and Gerhard F. Ecker. A binary QSAR model for classification of hERG potassium channel blockers. *Bioorganic & Medicinal Chemistry*, 16(7):4107–4119, April 2008.

[146] Vladimir Naumovich Vapnik. *The nature of statistical learning theory.* Springer, New York, 1995.

[147] V.N. Vapnik. An overview of statistical learning theory. *Neural Networks, IEEE Transactions on*, 10(5):988–999, September 1999.

[148] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, September 1995.

[149] Lian Yi Han, Chan Juan Zheng, Bin Xie, Jia Jia, Xiao Hua Ma, Feng Zhu, Hong Huang Lin, Xin Chen, and Yu Zong Chen. Support vector machines approach for predicting druggable proteins: recent progress in its exploration and investigation of its usefulness. *Drug Discovery Today*, 12(7–8):304–313, April 2007.

[150] J. Kamruzzaman and R.K. Begg. Support vector machines and other pattern recognition approaches to the diagnosis of cerebral palsy gait. *Biomedical Engineering, IEEE Transactions on*, 53(12):2479–2490, December 2006.

[151] William S Noble. What is a support vector machine? *Nat Biotech*, 24(12):1565–1567, December 2006.

[152] Christopher J.C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, June 1998.

[153] Support vector machine – wikipedia. http://de.wikipedia.org/wiki/ Support_Vector_Machine.

[154] John C. Platt. Fast training of support vector machines using sequential minimal optimization. In *Advances in kernel methods*, pages 185–208. MIT Press, 1999.

[155] E. Osuna, R. Freund, and F. Girosi. An improved training algorithm for support vector machines. In *Neural Networks for Signal Processing [1997] VII. Proceedings of the 1997 IEEE Workshop*, pages 276–285, September 1997.

[156] Karush–Kuhn–Tucker conditions - wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Karush%E2%80%93Kuhn%E2%80%93 Tucker_conditions.

[157] Craig L. Bruce, James L. Melville, Stephen D. Pickett, and Jonathan D. Hirst. Contemporary QSAR classifiers compared. *J. Chem. Inf. Model.*, 47(1):219–227, January 2007.

[158] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The WEKA data mining software: An update.

[159] The r project for statistical computing. http://www.r-project.org/.

[160] R. Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2012. ISBN 3-900051-07-0.

[161] Alexandros Karatzoglou, David Meyer, and Kurt Hornik. Support vector machines in r. *Journal of Statistical Software*, 15(i09).

[162] CRAN - package e1071. http://cran.r-project.org/web/packages/e1071/index.html.

[163] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm.

[164] Pai-Hsuen Chen, Rong-En Fan, and Chih-Jen Lin. Training support vector machines via SMO-Type decomposition methods. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Dough Tygar, Moshe Y. Vardi, Gerhard Weikum, Sanjay Jain, Hans Ulrich Simon, and Etsuji Tomita, editors, *Algorithmic Learning Theory*, volume 3734, pages 45–62. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.

[165] I. H Witten and Eibe Frank. *Data mining : practical machine learning tools and techniques*. Morgan Kaufman, Amsterdam; Boston, MA, 2005.

[166] Leo Breiman. *Classification and regression trees*. Chapman & Hall : ITP International Thomson Publishing, New York [etc.], 1984.

[167] Steven L. Salzberg. C4.5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993. *Machine Learning*, 16(3):235–240, September 1994.

[168] Usama M. Fayyad and Keki B. Irani. On the handling of continuous-valued attributes in decision tree generation. *Machine Learning*, 8(1):87–102, January 1992.

[169] Eric Deconinck, Menghui H. Zhang, Danny Coomans, and Yvan Vander Heyden. Classification tree models for the prediction of Blood-Brain barrier passage of drugs. *Journal of Chemical Information and Modeling*, 46(3):1410–1419, February 2006.

[170] Gini coefficient - wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Gini_coefficient.

[171] Information gain in decision trees - wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Information_gain_in_decision_trees.

[172] Claudia Andres and Michael C. Hutter. CNS permeability of drugs predicted by a decision tree. *QSAR & Combinatorial Science*, 25(4):305–309, 2006.

[173] Heping Zhang, Chang-Yung Yu, and Burton Singer. Cell and tumor classification using gene expression data: Construction of forests. *Proceedings of the National Academy of Sciences*, 100(7):4168–4172, April 2003.

[174] Anantha Prasad, Louis Iverson, and Andy Liaw. Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. *Ecosystems*, 9(2):181–199, March 2006.

[175] Thomas G. Dietterich. Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*, page 1–15. Springer-Verlag, 2000.

[176] Leo Breiman. Out-of-bag estimation. Technical report, 1996.

[177] Yoav Freund and Robert E. Schapire. A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence*, 14(5):771–780, 1999.

[178] Robert E. Schapire. A brief introduction to boosting. In *IJCAI*, pages 1401–1406, 1999.

[179] Richard Maclin and David Opitz. An empirical evaluation of bagging and boosting. In *In Proceedings of the Fourteenth National Conference on Artificial Intelligence*, page 546–551. AAAI Press, 1997.

[180] J. R. Quinlan. Bagging, boosting, and c4.5. In *In Proceedings of the Thirteenth National Conference on Artificial Intelligence*, page 725–730. AAAI Press, 1996.

[181] Leo Breiman. Population theory for boosting ensembles. *The Annals of Statistics*, 32(1):1–11, February 2003.

[182] Geoffrey I. Webb. MultiBoosting: a technique for combining boosting and wagging. *Mach. Learn.*, 40(2):159–196, August 2000.

[183] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, October 2001.

[184] Vladimir Svetnik, Andy Liaw, Christopher Tong, J. Christopher Culberson, Robert P. Sheridan, and Bradley P. Feuston. Random forest: A classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences*, 43(6):1947–1958, November 2003.

[185] Leo Breiman and Adele Cutler. Random forests. http://www.stat.berkeley.edu/~breiman/RandomForests/.

[186] Thomas Dietterich and Eun Bae Kong. Machine learning bias, statistical bias, and statistical variance of decision tree algorithms. Technical report, 1995.

[187] Vladimir Svetnik, Andy Liaw, Christopher Tong, and Ting Wang. Application of breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules. volume 3077 of *Lecture Notes in Computer Science*, pages 334–343. Springer Berlin / Heidelberg, 2004.

[188] Vladimir Svetnik, Ting Wang, Christopher Tong, Andy Liaw, Robert P. Sheridan, and Qinghua Song. Boosting: An ensemble learning tool for compound classification and QSAR modeling. *Journal of Chemical Information and Modeling*, 45(3):786–799, April 2005.

[189] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1):25, 2007.

[190] Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9(1):307, 2008.

[191] Kristin Nicodemus, James Malley, Carolin Strobl, and Andreas Ziegler. The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics*, 11(1):110, 2010.

[192] Weida Tong, Huixiao Hong, Hong Fang, Qian Xie, and Roger Perkins. Decision forest: Combining the predictions of multiple independent decision tree models. *Journal of Chemical Information and Computer Sciences*, 43(2):525–531, February 2003.

[193] Thomas M. Ehrman, David J. Barlow, and Peter J. Hylands. Virtual screening of chinese herbs with random forest. *Journal of Chemical Information and Modeling*, 47(2):264–278, January 2007.

[194] Robert Kirk DeLisle and Steven L. Dixon. Induction of decision trees via evolutionary programming. *Journal of Chemical Information and Computer Sciences*, 44(3):862–870, March 2004.

[195] Avrim L. Blum and Pat Langley. Selection of relevant features and examples in machine learning. *Relevance*, 97(1–2):245–271, December 1997.

[196] Andy Liaw and Matthew Wiener. Classification and regression by randomForest. *R News*, 2(3):18–22, 2002.

[197] S. Balakrishnama and A. Ganapathiraju. Linear discriminant analysis - a brief tutorial, March 1998.

[198] P. A. Lachenbruch and M. Goldstein. Discriminant analysis. *Biometrics*, 35(1):69–85, March 1979.

[199] Rajarshi Guha. Ph.D. thesis. http://rguha.net/writing/pub/thesis/thesis.html, 2005.

[200] A. Baldovin, W. Wen, D.L. Massart, and A. Turello. Regularised discriminant analysis (RDA) - modelling for the binary discrimination between pollution types. *Chemometrics and Intelligent Laboratory Systems*, 38(1):25–37, August 1997.

[201] W. Wu, Y. Mallet, B. Walczak, W. Penninckx, D.L. Massart, S. Heuerding, and F. Erni. Comparison of regularized discriminant analysis linear discriminant analysis and quadratic discriminant analysis applied to NIR data. *Analytica Chimica Acta*, 329(3):257–265, August 1996.

[202] Santosh Srivastava, Maya R. Gupta, and Béla A. Frigyik. Bayesian quadratic discriminant analysis. *Journal of Machine Learning Research*, 8:1277–1305, 2007.

[203] G.P. Moss, A.J. Shah, R.G. Adams, N. Davey, S.C. Wilkinson, W.J. Pugh, and Y. Sun. The application of discriminant analysis and machine learning methods as tools to identify and classify compounds with potential as transdermal enhancers. *European Journal of Pharmaceutical Sciences*, 45(1–2):116–127, January 2012.

[204] Mizanur R. Khondoker, Till T. Bachmann, Muriel Mewissen, Paul Dickinson, Bartosz Dobrzelecki, Colin J. Campbell, Andrew R. Mount, Anthony J. Walton, Jason Crain, Schulze Holger, Gerard Giraud, Alan J. Ross, Ilenia Ciani, Stuart W.J. Ember, Chaker Tlili, Jonathan G. Terry, Eilidh Grant, Nicola McDonnell, and Peter Ghazal. Multi-factorial analysis of class prediction error: estimating optimal number of biomarkers for various classification rules. *Journal of Bioinformatics and Computational Biology*, 08(06):945–965, December 2010.

[205] Aik Choon Tan, Daniel Q. Naiman, Lei Xu, Raimond L. Winslow, and Donald Geman. Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics*, 21(20):3896–3904, October 2005.

[206] Arja Asikainen, Mikko Kolehmainen, Juhani Ruuskanen, and Kari Tuppurainen. Structure-based classification of active and inactive estrogenic compounds by decision tree, LVQ and kNN methods. *Chemosphere*, 62(4):658–673, January 2006.

[207] C.W. Yap, Z.R. Li, and Y.Z. Chen. Quantitative structure–pharmacokinetic relationships for drug clearance by using statistical learning methods. *Journal of Molecular Graphics and Modelling*, 24(5):383–395, March 2006.

[208] Ahmed Shamsul Arefin, Carlos Riveros, Regina Berretta, and Pablo Moscato. GPU-FS-kNN: a software tool for fast and scalable kNN computation using GPUs. *PLoS ONE*, 7(8):e44000, August 2012.

[209] Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "Nearest neighbor" meaningful? In Catriel Beeri and Peter Buneman, editors, *Database Theory — ICDT'99*, volume 1540 of *Lecture Notes in Computer Science*, pages 217–235. Springer Berlin Heidelberg, January 1999.

[210] Dominik Kaiser, Barbara Zdrazil, and Gerhard F Ecker. Similarity-based descriptors (SIBAR)–a tool for safe exchange of chemical information? *Journal of computer-aided molecular design*, 19(9-10):687–692, October 2005. PMID: 16249834.

[211] Khac-Minh Thai and Gerhard F Ecker. Similarity-based SIBAR descriptors for classification of chemically diverse hERG blockers. *Molecular diversity*, 13(3):321–336, August 2009. PMID: 19219559.

[212] Tudor I. Oprea and Johan Gottfries. Chemography: The art of navigating in chemical space. http://pubs.acs.org/doi/abs/10.1021/cc0000388, February 2001.

[213] Rita Schwaha and Gerhard F. Ecker. Similarity based descriptors – useful for classification of substrates of the human multidrug transporter p-glycoprotein? *QSAR & Combinatorial Science*, 28(8):834–839, 2009.

[214] T I Oprea. Property distribution of drug-related chemical databases. *Journal of computer-aided molecular design*, 14(3):251–264, March 2000. PMID: 10756480.

[215] Johann Gasteiger and Mario Marsili. Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron*, 36(22):3219–3228, 1980.

[216] Alexandru T. Balaban. Highly discriminating distance-based topological index. *Chemical Physics Letters*, 89(5):399–404, July 1982.

[217] Paul Labute. A widely applicable set of descriptors. *Journal of Molecular Graphics and Modelling*, 18(4–5):464–477, 2000.

[218] Scott A. Wildman and Gordon M. Crippen. Prediction of physicochemical parameters by atomic contributions. *Journal of Chemical Information and Computer Sciences*, 39(5):868–873, August 1999.

[219] Robert S. Pearlman and K.M. Smith. Novel software tools for chemical diversity. *Perspectives in Drug Discovery and Design*, 9-11(0):339–353, January 1998.

[220] Harry Wiener. Structural determination of paraffin boiling points. *Journal of the American Chemical Society*, 69(1):17–20, January 1947.

[221] Lowell H. Hall and Lemont B. Kier. The molecular connectivity chi indexes and kappa shape indexes in structure-property modeling. In *Reviews in Computational Chemistry*, pages 367–422. John Wiley $0 Sons, Inc., 1991.

[222] QuaSAR-Descriptor. http://www.chemcomp.com/journal/descr.htm#KH.

[223] Liying Zhang, Hao Zhu, Tudor Oprea, Alexander Golbraikh, and Alexander Tropsha. QSAR modeling of the Blood–Brain barrier permeability for diverse organic compounds. *Pharmaceutical Research*, 25(8):1902–1914, 2008. 10.1007/s11095-008-9609-0.

[224] Khac-Minh Thai and Gerhard F. Ecker. Classification models for hERG inhibitors by counter-propagation neural networks. *Chemical Biology & Drug Design*, 72(4):279–289, 2008.

[225] Gilles Moreau and Pierre Broto. The autocorrelation of a topological structure: A new molecular descriptor. *Nouveau Journal de Chimie*, 6:359–360, 1980.

[226] Sebastian Schneegans. 2D- und 3D-Autokorrelationsdeskriptoren. Technical report, 2001.

[227] Boris Hollas. An analysis of the autocorrelation descriptor for molecules. *Journal of Mathematical Chemistry*, 33(2):91–101, February 2003.

[228] Driss Zakarya, Fathallah Tiyal, and Maurice Chastrette. Use of the multifunctional autocorrelation method to estimate molar volumes of alkanes and oxygenated compounds. comparison between components of autocorrelation vectors and topological indices. *Journal of Physical Organic Chemistry*, 6(10):574–582, 1993.

[229] Markus Wagener, Jens Sadowski, and Johann Gasteiger. Autocorrelation of molecular surface properties for modeling corticosteroid binding globulin and cytosolic ah receptor activity by neural networks. *Journal of the American Chemical Society*, 117(29):7769–7775, July 1995.

[230] ADRIANA.Code - calculation of molecular descriptors | inspiring chemical discovery. http://www.molecular-networks.com/products/adrianacode.

[231] Stefano Moro, Magdalena Bacilieri, Cristina Ferrari, and Giampiero Spalluto. Autocorrelation of molecular electrostatic potential surface properties combined with partial least squares analysis as alternative attractive tool to generate ligand-based 3D-QSARs. *Current drug discovery technologies*, 2(1):13–21, March 2005. PMID: 16472237.

[232] Maolin Wang, Kai Wang, Aixia Yan, and Changyuan Yu. Classification of HCV NS5B polymerase inhibitors using support vector machine. *International Journal of Molecular Sciences*, 13(4):4033–4047, March 2012.

[233] Manuel Pastor, Gabriele Cruciani, Iain McLay, Stephen Pickett, and Sergio Clementi. GRid-INdependent descriptors (GRIND): a novel class of alignment-independent three-dimensional molecular descriptors. *Journal of Medicinal Chemistry*, 43(17):3233–3243, August 2000.

[234] G. Cruciani, P. Crivori, P.-A. Carrupt, and B. Testa. Molecular fields in quantitative structure–permeation relationships: the VolSurf approach. *Journal of Molecular Structure: THEOCHEM*, 503(1–2):17–30, May 2000.

[235] Patrizia Crivori, Gabriele Cruciani, Pierre-Alain Carrupt, and Bernard Testa. Predicting Blood-Brain barrier permeation from three-dimensional molecular structure. *Journal of Medicinal Chemistry*, 43(11):2204–2216, May 2000.

[236] Giuseppe Ermondi, Giulia Caron, Isela Garcia Pintos, Michela Gerbaldo, Manuel Pérez, Daniel I. Pérez, Zoila Gándara, Ana Martínez, Generosa Gómez, and Yagamare Fall. An application of two MIFs-based tools (volsurf+ and pentacle) to binary QSAR: the case of a palinurin-related data set of non-ATP competitive glycogen synthase kinase 3β (GSK-3β) inhibitors. *European Journal of Medicinal Chemistry*, 46(3):860–869, March 2011.

[237] Dong Dong and Baojian Wu. In silico modeling of UDP-glucuronosyltransferase 1A10 substrates using the volsurf approach. *Journal of Pharmaceutical Sciences*, 101(9):3531–3539, 2012.

[238] N. S. Hari Narayana Moorthy, Maria J. Ramos, and Pedro A. Fernandes. Comparative structural analysis of α-glucosidase inhibitors on difference species: A computational study. *Archiv der Pharmazie*, 345(4):265–274, 2012.

[239] Luciana Scotti, Elizabeth Igne Ferreira, Marcelo Sobral da Silva, and Marcus Tullius Scotti. Chemometric studies on natural products as potential inhibitors of the NADH oxidase from trypanosoma cruzi using the VolSurf approach. *Molecules*, 15(10):7363–7377, 2010.

[240] Brian B. Masek, Arshad Merchant, and James B. Matthew. Molecular shape comparison of angiotensin II receptor antagonists. *Journal of Medicinal Chemistry*, 36(9):1230–1238, April 1993.

[241] J. A. Grant, M. A. Gallardo, and B. T. Pickup. A fast method of molecular shape comparison: A simple application of a gaussian description of molecular shape. *Journal of Computational Chemistry*, 17(14):1653–1666, 1996.

[242] Michael L. Connolly. Computation of molecular volume. *Journal of the American Chemical Society*, 107(5):1118–1124, March 1985.

[243] Thomas S. Rush, J. Andrew Grant, Lidia Mosyak, and Anthony Nicholls. A shape-based 3-d scaffold hopping method and its application to a bacterial Protein-Protein interaction. *Journal of Medicinal Chemistry*, 48(5):1489–1495, February 2005.

[244] Paul C. D. Hawkins, A. Geoffrey Skillman, and Anthony Nicholls. Comparison of shape-matching and docking as virtual screening tools. *Journal of Medicinal Chemistry*, 50(1):74–82, December 2006.

[245] Robert Sheridan, Georgia McGaughey, and Wendy Cornell. Multiple protein structures and multiple ligands: effects on the apparent goodness of virtual screening results. *Journal of Computer-Aided Molecular Design*, 22(3):257–265, March 2008.

[246] Jennifer Venhorst, Sara Nuñez, Jan Willem Terpstra, and Chris G. Kruse. Assessment of scaffold hopping efficiency by use of molecular interaction fingerprints. *Journal of Medicinal Chemistry*, 51(11):3222–3229, May 2008.

[247] Imran S Haque and Vijay S Pande. PAPER–accelerating parallel evaluations of ROCS. *Journal of computational chemistry*, 31(1):117–132, January 2010. PMID: 19421991.

[248] OMEGA | OpenEye scientific software. http://www.eyesopen.com/omega.

[249] Johannes Kirchmair, Simona Distinto, Patrick Markt, Daniela Schuster, Gudrun M. Spitzer, Klaus R. Liedl, and Gerhard Wolber. How to optimize shape-based virtual screening: Choosing the right query and including chemical information. *Journal of Chemical Information and Modeling*, 49(3):678–692, March 2009.

[250] MOE: molecular operating environment. http://www.chemcomp.com/ software.htm.

[251] Joseph L. Durant, Burton A. Leland, Douglas R. Henry, and James G. Nourse. Reoptimization of MDL keys for use in drug discovery. *Journal of Chemical Information and Computer Sciences*, 42(6):1273–1280, September 2002.

[252] Todd Ewing, J Christian Baber, and Miklos Feher. Novel 2D fingerprints for ligand-based virtual screening. *Journal of chemical information and modeling*, 46(6):2423–2431, December 2006. PMID: 17125184.

[253] Yuan Wang and Jürgen Bajorath. Bit silencing in fingerprints enables the derivation of compound class-directed similarity metrics. *Journal of Chemical Information and Modeling*, 48(9):1754–1759, August 2008.

[254] CRAN - package MASS. http://cran.r-project.org/web/packages/MASS/index.html.

[255] CRAN - package knncat. http://www.icesi.edu.co/CRAN/web/packages/knncat/index.html.

[256] Ryoka Systems Inc. AutoQSAR: SVL exchange - an SVL code exchange site for the MOE user community. http://svl.chemcomp.com/.

[257] Alexander Hillebrecht and Gerhard Klebe. Use of 3D QSAR models for database screening: A feasibility study. *Journal of Chemical Information and Modeling*, 48(2):384–396, January 2008.

[258] Hao Zhu, Alexander Tropsha, Denis Fourches, Alexandre Varnek, Ester Papa, Paola Gramatica, Tomas Öberg, Phuong Dao, Artem Cherkasov, and Igor V. Tetko. Combinatorial QSAR modeling of chemical toxicants tested against tetrahymena pyriformis. *Journal of Chemical Information and Modeling*, 48(4):766–784, March 2008.

[259] ROCS - rapid overlay of chemical structures. *Version 2.2*, 2006.

[260] Rita Schwaha and Gerhard F Ecker. Use of shape similarities for the classification of p-glycoprotein substrates and nonsubstrates. *Future medicinal chemistry*, 3(9):1117–1128, July 2011. PMID: 21806376.

[261] Tomohiro Sato, Hitomi Yuki, Daisuke Takaya, Shunta Sasaki, Akiko Tanaka, and Teruki Honma. Application of support vector machine to three-dimensional shape-based virtual screening using comprehensive three-dimensional molecular shape overlay with known inhibitors. *Journal of Chemical Information and Modeling*, 52(4):1015–1026, March 2012.

[262] Guoping Hu, Guanglin Kuang, Wen Xiao, Weihua Li, Guixia Liu, and Yun Tang. Performance evaluation of 2D fingerprint and 3D shape similarity methods in virtual screening. *Journal of Chemical Information and Modeling*, 52(5):1103–1113, May 2012.

[263] Josefin Larsson, Johan Gottfries, Lars Bohlin, and Anders Backlund. Expanding the ChemGPS chemical space with natural products. *Journal of natural products*, 68(7):985–991, July 2005. PMID: 16038536.

Appendix A

Dataset

Table A.1: Compounds of the training set

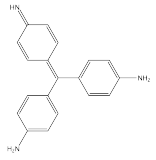
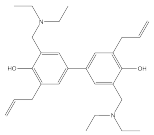
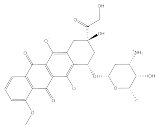
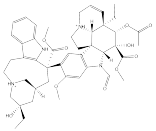
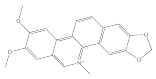
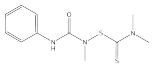
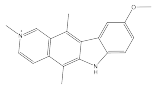
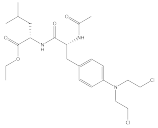
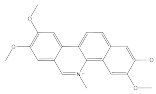
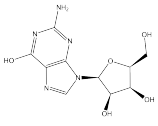
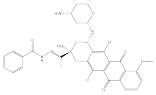
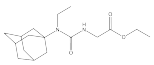
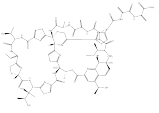
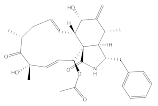
Structures	Name	sub.	Structures	Name	sub.
	NSC10460	1		NSC6386	0
	NSC123127	1		NSC67574	0
	NSC146397	1		NSC161128	0
	NSC155693	1		NSC167780	0
	NSC157995	1		NSC19994	0
	NSC164011	1		NSC208912	0
	NSC170365	1		NSC209835	0

Table A.1: Compounds of the training set

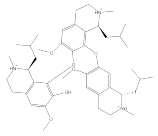
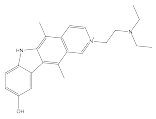
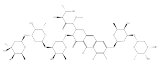
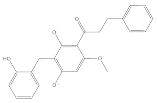
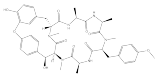
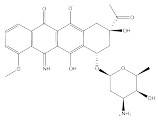
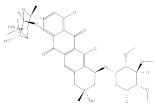
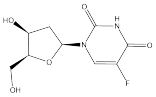
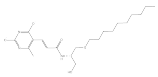
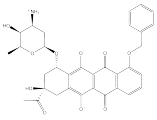
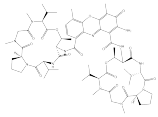
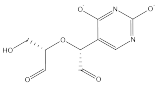
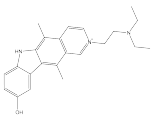
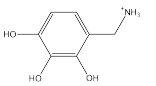
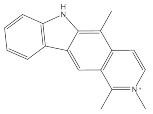
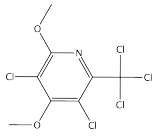
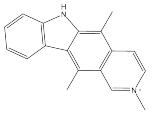
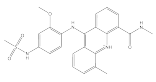
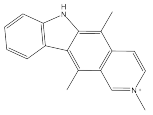
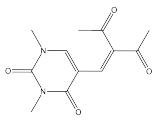
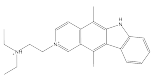
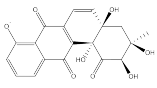
Structures	Name	sub.	Structures	Name	sub.
	NSC21075	1		NSC227279	0
	NSC24559	1		NSC241906	0
	NSC259968	1		NSC254681	0
	NSC265450	1		NSC27640	0
	NSC265473	1		NSC286628	0
	NSC3053	1		NSC291643	0
	NSC311152	1		NSC329097	0
	NSC336003	1		NSC338720	0
	NSC351710	1		NSC343499	0
	NSC352299	1		NSC353882	0
	NSC353076	1		NSC368697	0

Table A.1: Compounds of the training set

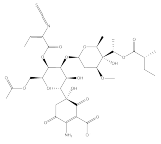
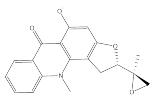
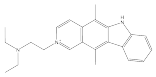
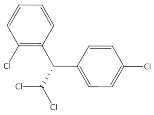
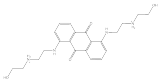
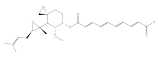
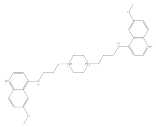
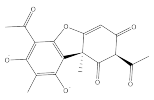
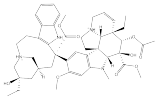
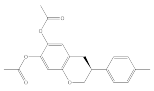
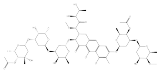
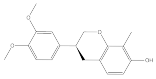
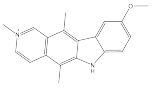
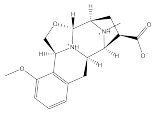
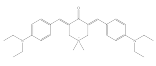
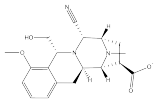
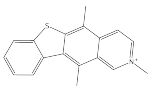
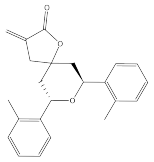
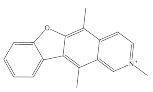
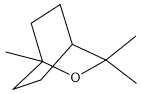
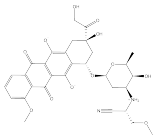
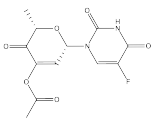
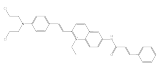
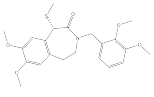
Structures	Name	sub.	Structures	Name	sub.
	NSC356207	1		NSC383031	0
	NSC359449	1		NSC38721	0
	NSC363997	1		NSC58368	0
	NSC365360	1		NSC5890	0
	NSC49842	1		NSC600288	0
	NSC58514	1		NSC600291	0
	NSC627505	1		NSC601422	0
	NSC634791	1		NSC607097	0
	NSC638066	1		NSC617131	0
	NSC638788	1		NSC6171	0
	NSC639659	1		NSC618093	0
	NSC640085	1		NSC619859	0

Table A.1: Compounds of the training set

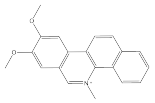
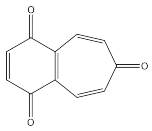
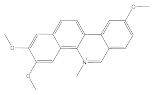
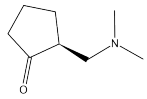
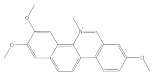
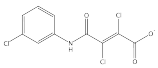
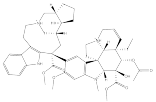
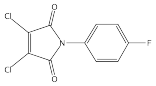
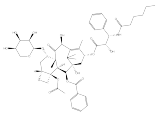
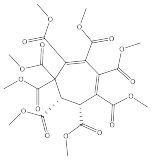
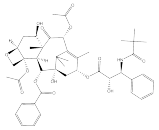
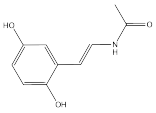
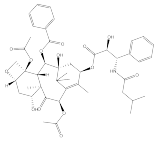
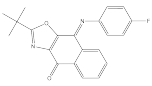
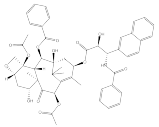
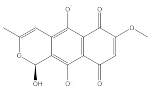
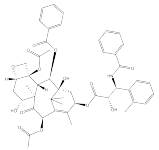
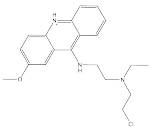
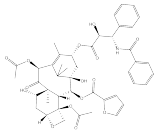
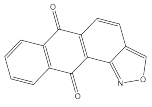
Structures	Name	sub.	Structures	Name	sub.
	NSC645301	1		NSC620056	0
	NSC645305	1		NSC621888	0
	NSC645306	1		NSC622381	0
	NSC651727	1		NSC622384	0
	NSC658831	1		NSC625133	0
	NSC664402	1		NSC625301	0
	NSC664404	1		NSC626030	0
	NSC666608	1		NSC626482	0
	NSC671870	1		NSC628114	0
	NSC673187	1		NSC628939	0

Table A.1: Compounds of the training set

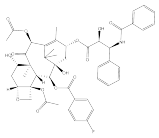
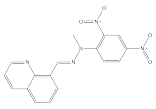
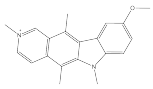
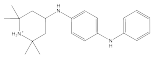
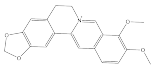
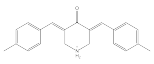
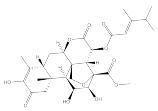
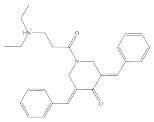
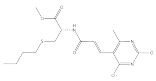
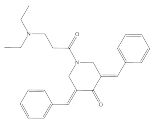
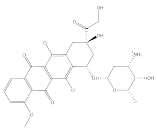
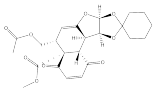
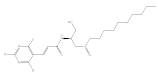
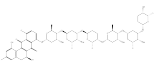
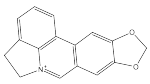
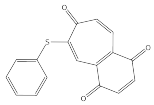
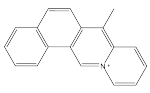
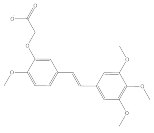
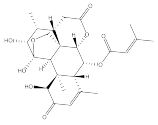
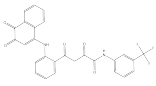
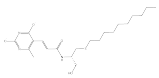
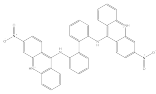
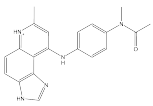
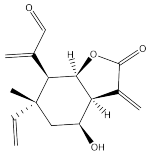
Structures	Name	sub.	Structures	Name	sub.
	NSC673188	1		NSC630684	0
	NSC155694	1		NSC632536	0
	NSC163088	1		NSC632839	0
	NSC165563	1		NSC634785	0
	NSC201241	1		NSC634786	0
	NSC256942	1		NSC637399	0
	NSC266763	1		NSC639187	0
	NSC270693	1		NSC640192	0
	NSC28002	1		NSC643813	0
	NSC290494	1		NSC644751	0
	NSC305458	1		NSC645158	0
	NSC305884	1		NSC645991	0

Table A.1: Compounds of the training set

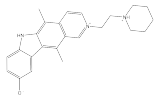
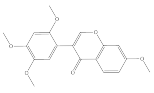
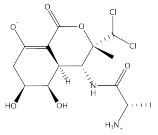
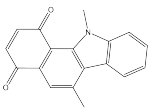
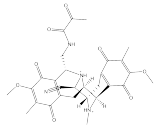
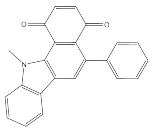
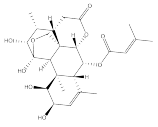
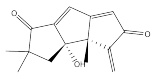
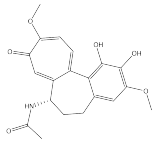
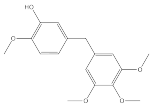
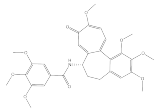
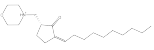
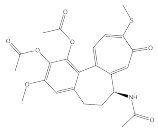
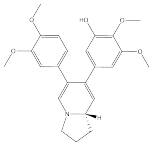
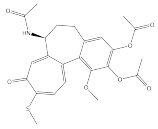
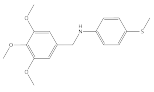
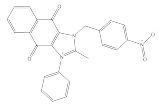
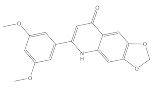
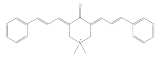
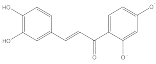
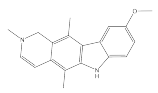
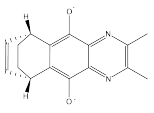
Structures	Name	sub.	Structures	Name	sub.
	NSC311153	1		NSC646923	0
	NSC325014	1		NSC648147	0
	NSC325663	1		NSC648150	0
	NSC341651	1		NSC648322	0
	NSC354975	1		NSC648581	0
	NSC355256	1		NSC649910	0
	NSC374979	1		NSC650396	0
	NSC374980	1		NSC650772	0
	NSC52745	1		NSC652112	0
	NSC636679	1		NSC652892	0
	NSC637651	1		NSC658387	0

Table A.1: Compounds of the training set

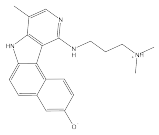
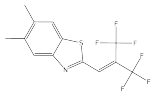
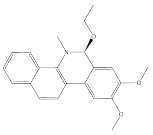
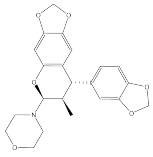
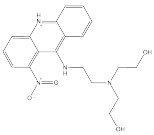
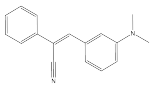
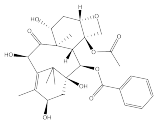
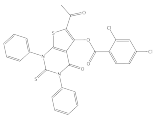
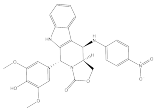
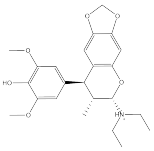
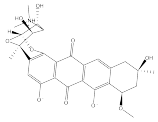
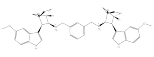
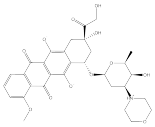
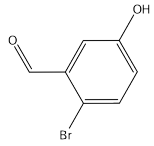
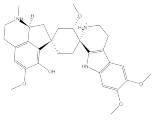
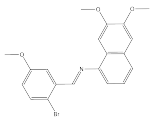
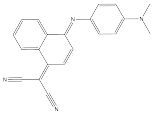
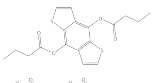
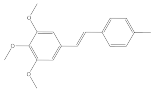
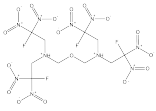
Structures	Name	sub.	Structures	Name	sub.
	NSC645008	1		NSC664311	0
	NSC645302	1		NSC666222	0
	NSC645806	1		NSC667251	0
	NSC656178	1		NSC671136	0
	NSC668380	1		NSC671170	0
	NSC269148	0		NSC674066	0
	NSC354646	0		NSC680715	0
	NSC626578	0		NSC680717	0
	NSC627777	0		NSC682994	0
	NSC638485	0		NSC683257	0

Table A.1: Compounds of the training set

Structures	Name	sub.	Structures	Name	sub.

Table A.1: Compounds of the training set.

Table A.2: Compounds of the test set

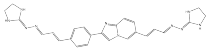
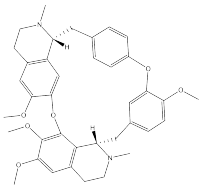
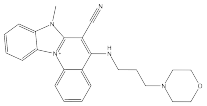
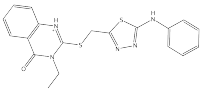
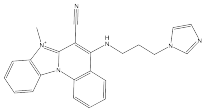
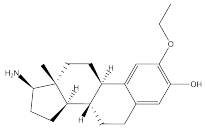
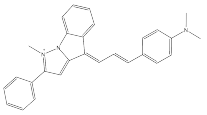
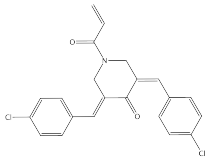
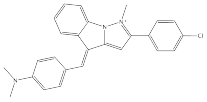
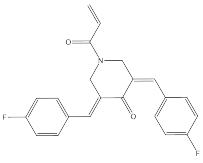
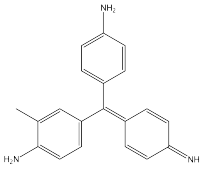
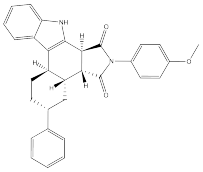
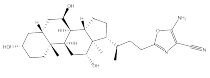
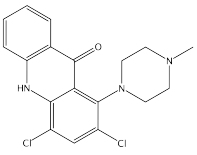
Structure	Name	sub.	Structure	Name	sub.
	NSC690242	1		NSC77037	0
	NSC694262	1		NSC686368	0
	NSC694268	1		NSC687454	0
	NSC699477	1		NSC687849	0
	NSC699479	1		NSC687850	0
	NSC93739	1		NSC689131	0
	NSC685302	1		NSC690568	0

Table A.2: Compounds of the test set

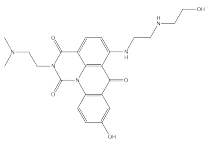
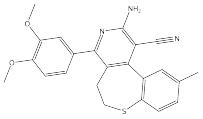
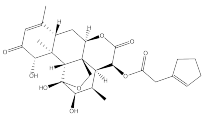
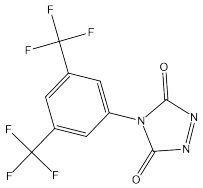
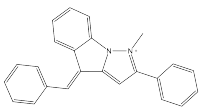
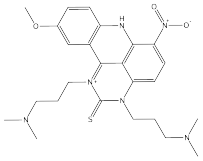
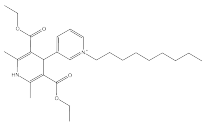
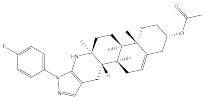
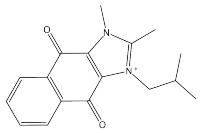
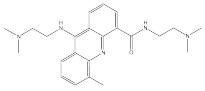
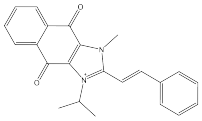
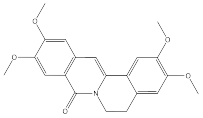
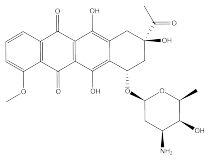
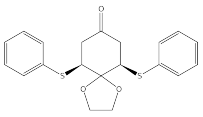
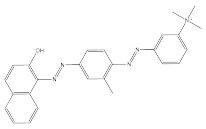
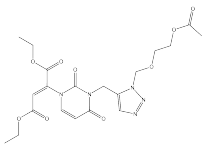
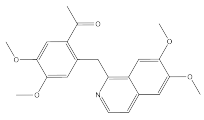
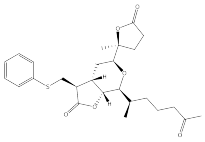
Structure	Name	sub.	Structure	Name	sub.
	NSC693120	1		NSC691562	0
	NSC693539	1		NSC691612	0
	NSC693575	1		NSC691845	0
	NSC695636	1		NSC692655	0
	NSC80466	1		NSC692738	0
	NSC80469	1		NSC693145	0
	NSC82151	1		NSC693224	0
	NSC9609	1		NSC693992	0
	NSC98542	1		NSC697266	0

Table A.2: Compounds of the test set

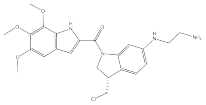
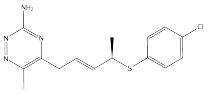
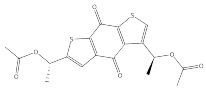
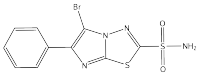
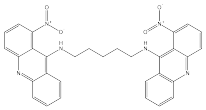
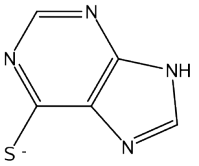
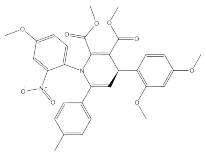
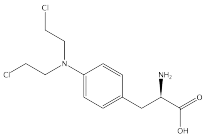
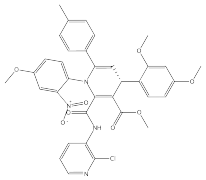
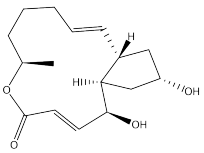
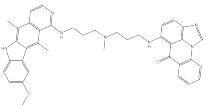
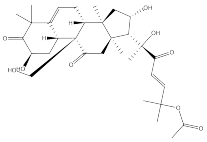
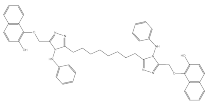
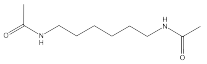
Structure	Name	sub.	Structure	Name	sub.
	NSC688304	0		NSC698023	0
	NSC690435	0		NSC698131	0
	NSC690635	0		NSC755	0
	NSC692574	0		NSC8806	0
	NSC692576	0		NSC89671	0
	NSC695938	0		NSC94743	0
	NSC697168	0		NSC95580	0

Table A.2: Compounds of the test set

Table A.3: Compounds of reference set A

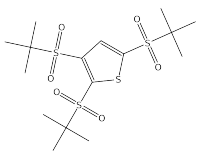
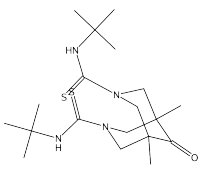
Structure	Name	Structure	Name
	20754		335663

Table A.3: Compounds of reference set A

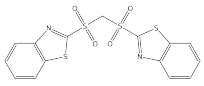
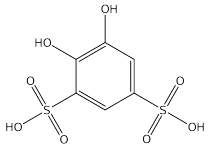
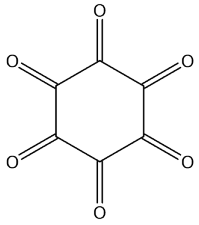
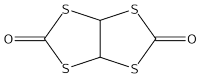
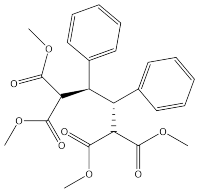
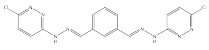
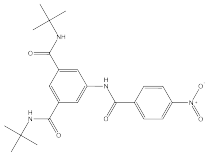
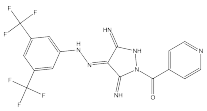
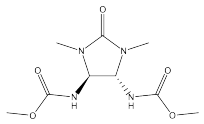
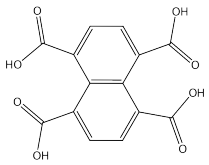
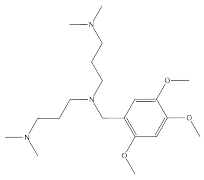
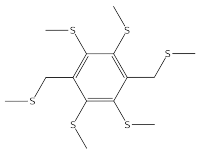
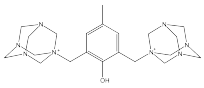
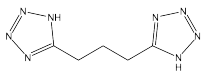
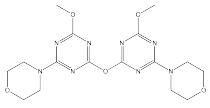
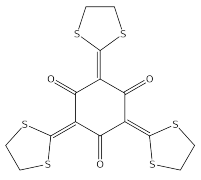
Structure	Name	Structure	Name
	21201		342788
	30731		414395
	36644		419875
	38964		490486
	42235		496329
	45652		507489
	58338		509915
	114608		537902

Table A.3: Compounds of reference set A

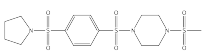
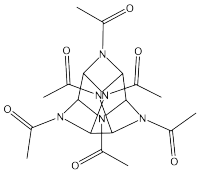
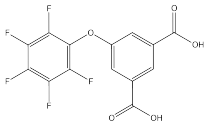
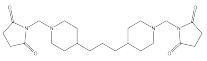
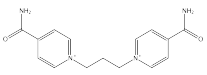
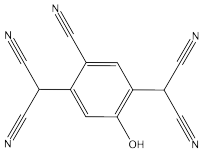
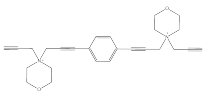
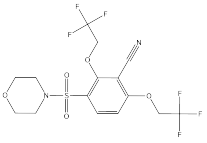
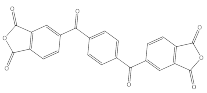
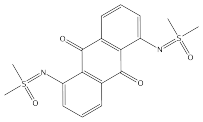
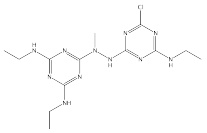
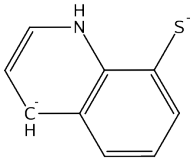
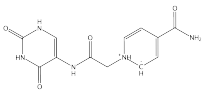
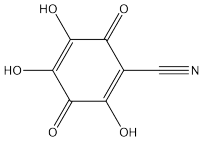
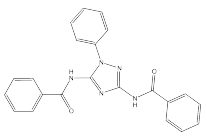
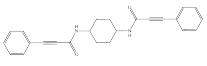
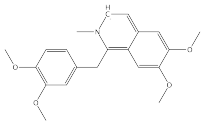
Structure	Name	Structure	Name
	152313		538239
	173764		551265
	173987		564582
	174595		567010
	175638		569419
	181022		592738
	214673		603322
	214675		603428
	235804		603587

Table A.3: Compounds of reference set A

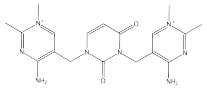
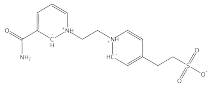
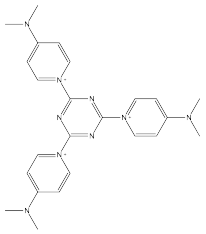
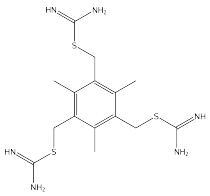
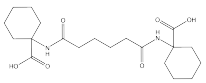
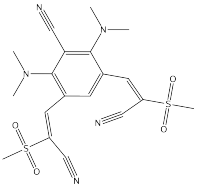
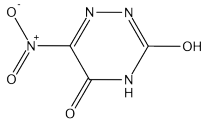
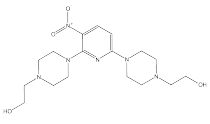
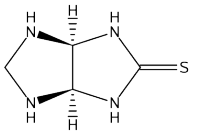
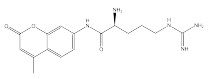
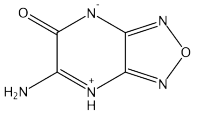
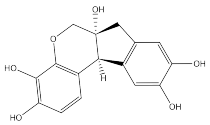
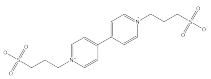
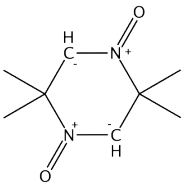
Structure	Name	Structure	Name
	237992		603679
	238465		627486
	239978		632082
	315918		632895
	318565		650172
	319278		654160
	335215		654199

Table A.3: Compounds of reference set A

Table A.4: Compounds of reference set B

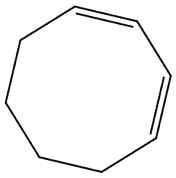
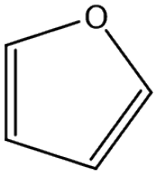
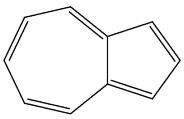
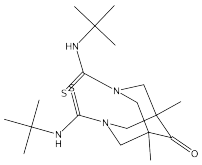
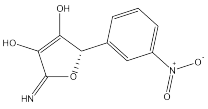
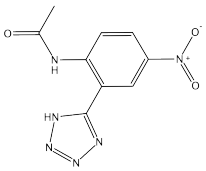
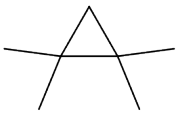
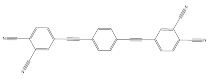
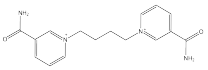
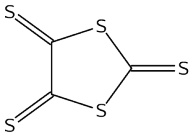
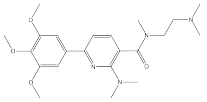
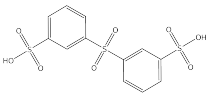
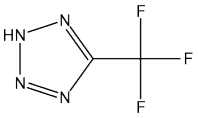
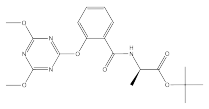
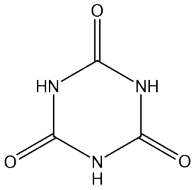
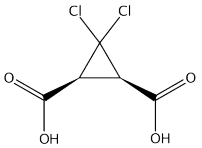
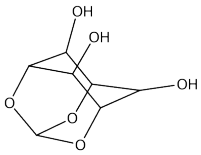
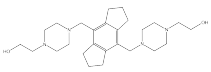
Structure	Name	Structure	Name
	3949		331470
	3998		335663
	4048		340555
	4164		343894
	6507		414406
	8445		427870
	23487		485054
	30372		490119
	30948		496121

Table A.4: Compounds of reference set B

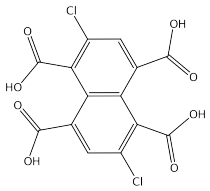
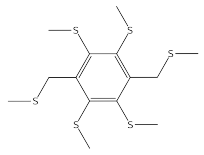
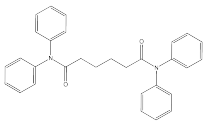
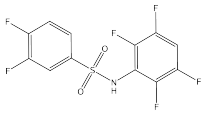
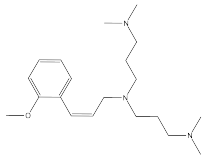
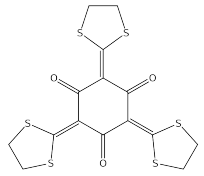
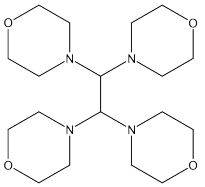
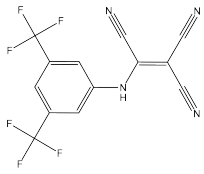
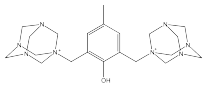
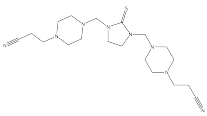
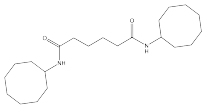
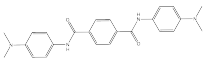
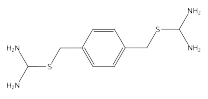
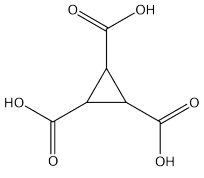
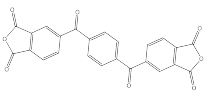
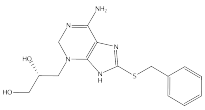
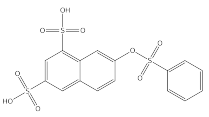
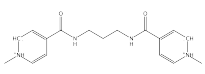
Structure	Name	Structure	Name
	38793		507489
	39233		526666
	45315		537902
	47991		555218
	58338		571379
	76554		582941
	125818		587558
	175638		613022
	196698		613630

Table A.4: Compounds of reference set B

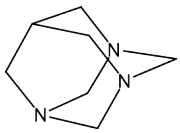
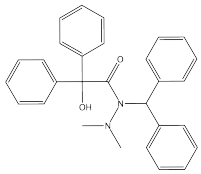
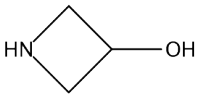
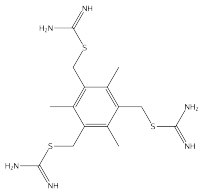
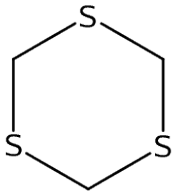
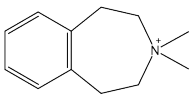
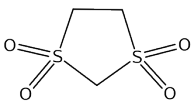
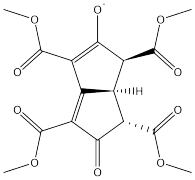
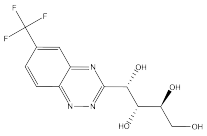
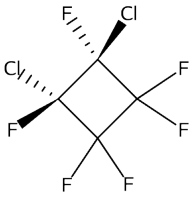
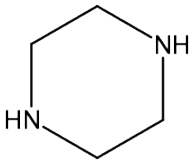
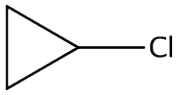
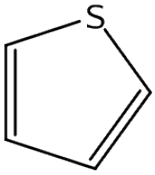
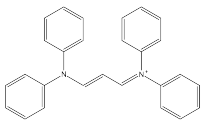
Structure	Name	Structure	Name
	196932		616481
	237335		627486
	316225		628659
	322895		629234
	331326		629415
	331337		629528
	331376		660615

Table A.5: Compounds of reference set C

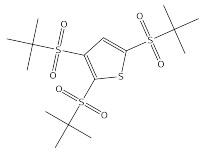
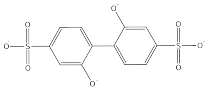
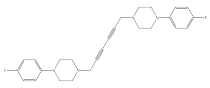
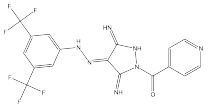
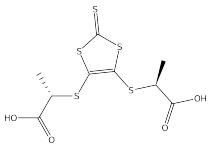
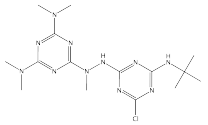
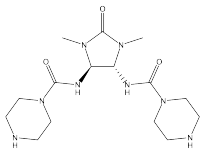
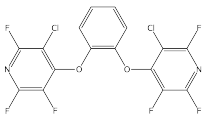
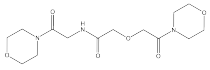
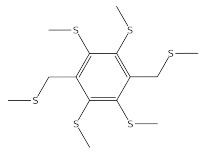
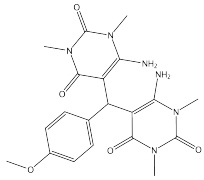
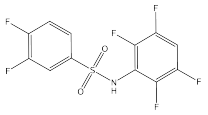
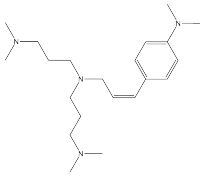
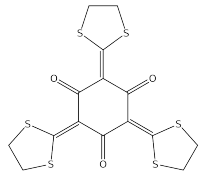
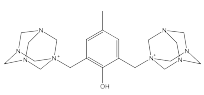
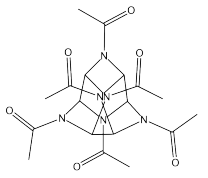
Structure	Name	Structure	Name
	20754		440947
	32635		490486
	33773		499368
	33950		500138
	38282		507489
	41789		526666
	45140		537902
	58338		538239

Table A.5: Compounds of reference set C

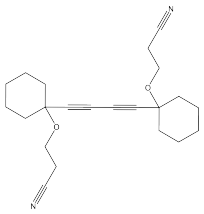
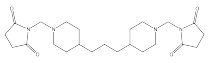
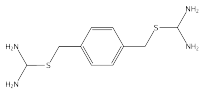
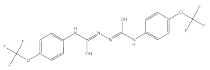
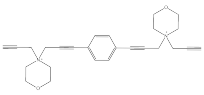
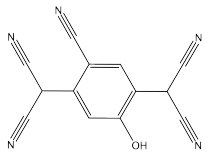
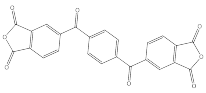
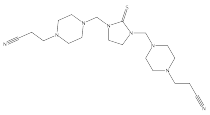
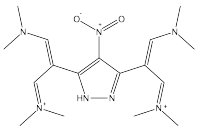
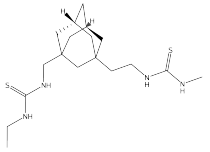
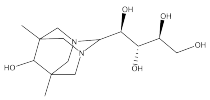
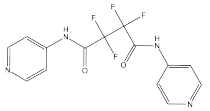
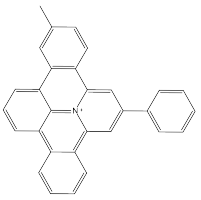
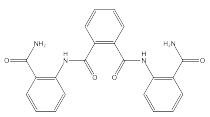
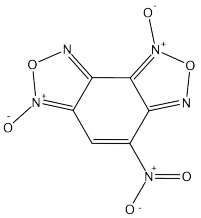
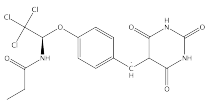
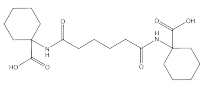
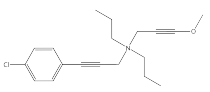
Structure	Name	Structure	Name
	92506		551265
	125818		556870
	174595		564582
	175638		571379
	175689		572522
	178264		582945
	238484		583252
	238942		583668
	239978		589743

Table A.5: Compounds of reference set C

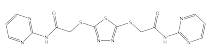
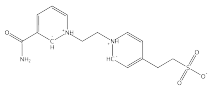
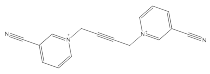
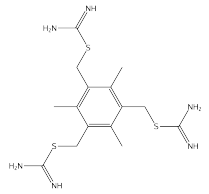
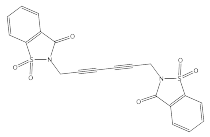
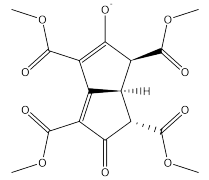
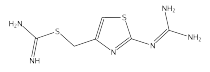
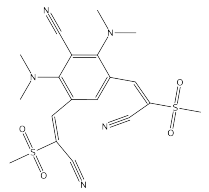
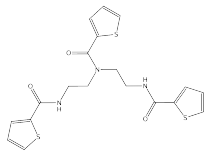
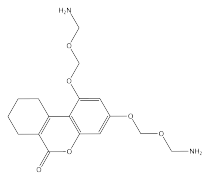
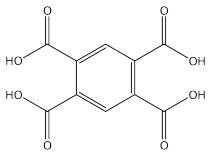
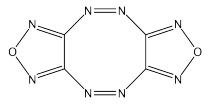
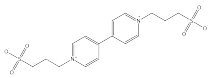
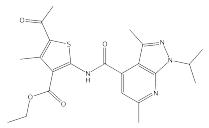
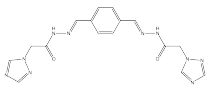
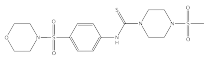
Structure	Name	Structure	Name
	241013		603679
	315654		627486
	320577		629234
	321062		632082
	325364		650845
	331463		651248
	335215		655728
	409960		656430

Table A.5: Compounds of reference set C

Table A.6: Compounds of reference set D

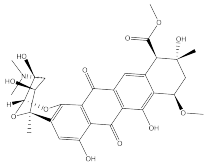
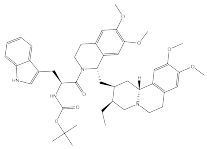
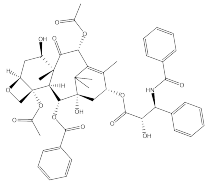
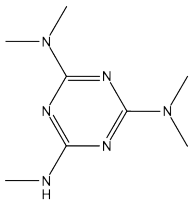
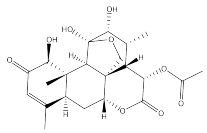
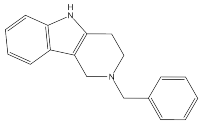
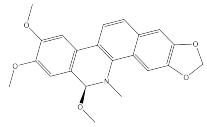
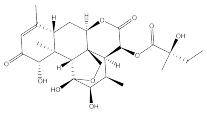
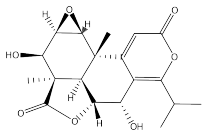
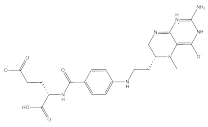
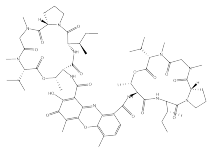
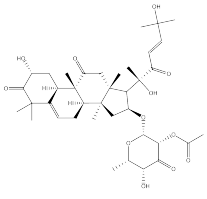
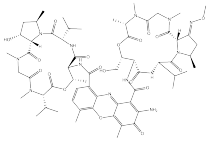
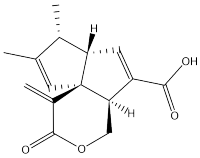
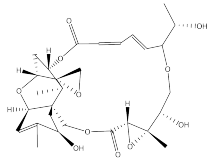
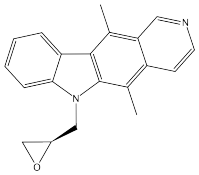
Structure	Name	sub.	Structure	Name	sub.
	NSC102815	1		NSC109350	0
	NSC125973	1		NSC118742	0
	NSC126765	1		NSC122301	0
	NSC146396	1		NSC132791	0
	NSC211500	1		NSC139490	0
	NSC237106	1		NSC144153	0
	NSC237671	1		NSC145150	0
	NSC269756	1		NSC152731	0

Table A.6: Compounds of reference set D

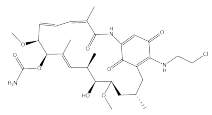
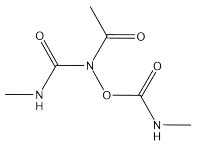
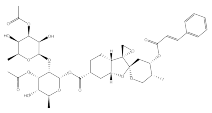
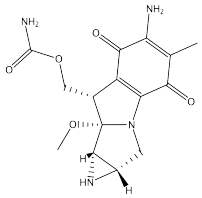
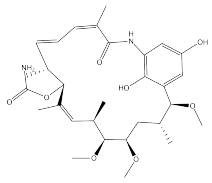
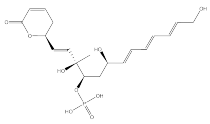
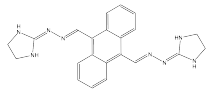
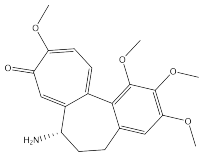
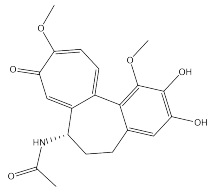
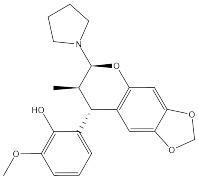
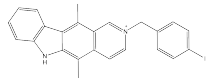
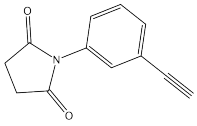
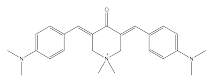
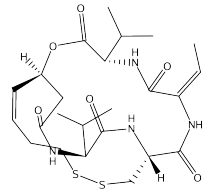
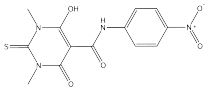
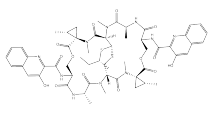
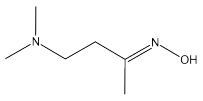
Structure	Name	sub.	Structure	Name	sub.
	NSC320877	1		NSC253272	0
	NSC328426	1		NSC26980	0
	NSC330500	1		NSC339638	0
	NSC337766	1		NSC36354	0
	NSC354974	1		NSC375503	0
	NSC604574	1		NSC634182	0
	NSC618757	1		NSC643828	0
	NSC630176	1		NSC645838	0
	NSC630678	1		NSC651590	0

Table A.6: Compounds of reference set D

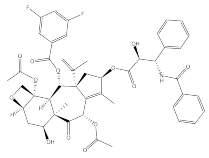
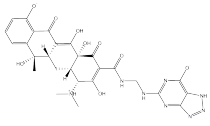
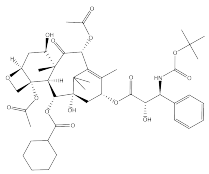
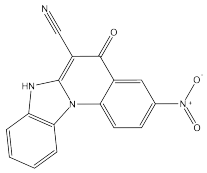
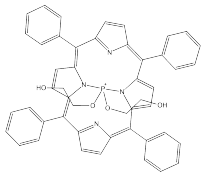
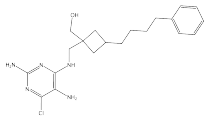
Structure	Name	sub.	Structure	Name	sub.
	NSC661746	1		NSC67586	0
	NSC671867	1		NSC691216	0
	NSC682066	1		NSC695805	0

Table A.6: Compounds of reference set D

Cross-validated results

SMO-Polykernel	10CV A	A1	A0	Pr1	Pr0
3DAuto RefA	71,53	66,07	76,40	71,36	71,67
3DAuto RefB	76,03	74,49	77,40	74,58	77,32
3DAuto RefC	72,33	71,12	73,40	70,41	74,07
3DAuto RefD	73,91	77,46	70,75	70,15	77,96
pure 3DAuto	76,49	76,48	76,50	74,28	78,56
VSA RefA	76,61	75,11	78,00	75,85	77,30
VSA RefB	76,09	75,00	77,10	75,08	77,02
VSA RefC	75,05	74,13	75,90	73,89	76,13
VSA RefD	77,43	75,83	78,88	76,36	78,39
pure VSA	80,53	78,33	82,50	80,11	80,88
2D RefA	75,78	71,96	79,30	76,18	75,45
2D RefB	79,79	83,26	76,60	76,60	83,26
2D RefC	74,95	76,74	73,30	72,56	77,40
2D RefD	79,28	78,61	79,88	77,85	80,58
pure 2D	83,42	85,00	82,00	80,95	85,86

Table A.7: Cross-validated results of support vector machine of study 1 - Polykernel

SMO-RBF-Kernel	10CV A	A1	A0	Pr1	Pr0
3DAuto RefA	72,49	73,48	67,87	69,72	71,77
3DAuto RefB	73,76	49,50	70,40	69,98	77,88
3DAuto RefC	72,75	72,58	72,90	70,45	74,92
3DAuto RefD	74,24	77,89	71,00	70,45	78,34
pure 3Dauto	80,33	77,32	83,00	80,15	80,48
VSA RefA	74,38	71,52	77,00	74,10	74,61
VSA RefB	74,22	73,48	74,90	72,92	75,43
VSA RefC	74,32	70,11	78,20	0,75	0,74
VSA RefD	79,28	80,00	78,63	77,11	81,37
pure VSA	80,92	82,36	79,63	78,44	83,38
2D RefA	78,18	80,87	75,70	75,38	81,14
2D RefB	74,22	76,63	72,00	71,57	77,01
2D RefC	73,49	75,65	71,50	70,95	76,14
2D RefD	76,18	72,64	79,38	76,02	76,32
pure 2D	82,43	82,36	82,50	80,90	83,86

Table A.8: Cross-validated results of support vector machine of study 1 - RBF Kernel

BQSAR- 15 Comp.	LOO XA	XA1	XA0	XPr1	XPr0	Comp. Used	Descr.
3DAuto RefA	75,13	62,92	86,00	80,00	72,27	15	50
3DAuto RefB	77,78	67,42	87,00	82,19	75,00	14	50
3DAuto RefC	80,95	69,66	91,00	87,32	77,12	15	50
3DAutoRefD	78,15	73,24	82,50	78,79	77,65	15	13
pure 3DAuto	85,43	76,06	93,75	91,53	81,52	15	84
VSA RefA	78,65	66,30	90,00	85,92	74,38	14	50
VSA RefB	77,60	70,65	84,00	80,25	75,68	10	50
VSA RefC	79,17	68,48	89,00	85,14	75,42	15	50
VSARefD	79,61	68,06	90,00	85,96	75,79	**12**	**40**
pure VSA	79,61	65,28	92,50	88,68	74,75	**14**	**32**
ADME RefA	77,60	78,26	77,00	75,79	79,38	12	50
ADME RefB	75,52	76,09	75,00	73,68	77,32	15	50
ADME RefC	77,08	76,09	78,00	76,09	78,00	11	50
ADMERefD	84,21	75,00	92,50	90,00	80,43	**15**	**40**
pure ADME	80,92	70,83	90,00	86,44	77,42	12	88

Table A.9: Cross-validated results of binary QSAR of study 1

Table A.10: Cross-validated results of binary QSAR of study 2

Crossval.	15 Comp. BQSAR Descriptors	A	A1	A0	Pr1	Pr0
RefA	ColorScore	80,95	72,46	88,46	84,75	78,41
	ComboScore	81,63	72,46	89,74	86,21	78,65
	Overlap	76,87	62,32	89,74	84,31	72,92
	ScaledColor	78,91	66,67	89,74	85,19	75,27
	ShapeTanimoto	78,91	68,12	88,46	83,93	75,82
	Tversky.d	79,59	73,91	84,62	80,95	78,57
	Tversky.q	82,99	72,46	92,31	89,29	79,12
RefB	ColorScore	76,87	63,77	88,46	83,02	73,40
	ComboScore	78,91	65,22	91,03	86,54	74,74
	Overlap	80,95	68,12	92,31	88,68	76,60
	ScaledColor	78,23	68,12	87,18	82,46	75,56
	ShapeTanimoto	85,03	73,91	94,87	92,73	80,43
	Tversky.d	72,79	63,77	80,77	74,58	71,59
	Tversky.q	86,39	75,36	96,15	94,55	81,52
RefC	ColorScore	76,19	63,77	87,18	81,48	73,12
	ComboScore	85,03	72,46	96,15	94,34	79,79
	Overlap	79,59	69,57	88,46	84,21	76,67
	ScaledColor	80,27	66,67	92,31	88,46	75,79
	ShapeTanimoto	81,63	72,46	89,74	86,21	78,65
	Tversky.d	75,51	62,32	87,18	81,13	72,34
	Tversky.q	85,71	73,91	96,15	94,44	80,65
RefD	ColorScore	80,27	65,22	93,59	90,00	75,26
	ComboScore	83,67	68,12	97,44	95,92	77,55
	Overlap	78,91	65,22	91,03	86,54	74,74
	ScaledColor	80,95	66,67	93,59	90,20	76,04
	ShapeTanimoto	82,99	72,46	92,31	89,29	79,12
	Tversky.d	78,23	69,57	85,90	81,36	76,14
	Tversky.q	75,51	62,32	87,18	81,13	72,34
	vSURF	77,40	63,24	89,74	84,31	73,68
	vSURF	78,77	70,59	85,90	81,36	77,01
	vSURF	74,66	67,65	80,77	75,41	74,12

Table A.10: Cross-validated results of binary QSAR of study 2

Crossval. 15 Comp. BQSAR Descriptors	A	A1	A0	Pr1	Pr0
vSURF	80,14	70,59	88,46	84,21	77,53
pure vSURF	76,71	60,29	91,03	85,42	72,45
VSA	84,25	73,53	93,59	90,91	80,22
VSA	76,71	64,71	87,18	81,48	73,91
VSA	82,19	67,65	94,87	92,00	77,08
VSA	80,14	64,71	93,59	89,80	75,26
pure VSA	78,77	63,24	92,31	87,76	74,23

Table A.10: Cross-validated results of binary QSAR of study 2

Table A.11: Cross-validated results of random forest

Crossval. Random Forest	15 Comp. Descriptors	A	A1	A0	Pr1	Pr0
RefA	ColorScore	73,52	67,81	80,00	73,76	73,46
	ComboScore	76,10	73,83	80,69	76,33	77,19
	Overlap	79,67	84,83	75,00	74,80	83,78
	ScaledColor	70,81	64,71	77,50	70,96	70,78
	ShapeTanimoto	76,38	73,05	80,42	79,53	75,52
	Tversky.d	75,57	75,61	75,56	74,66	76,81
	Tversky.q	76,86	78,15	75,28	73,90	79,22
RefB	ColorScore	75,00	69,32	80,28	73,63	75,20
	ComboScore	71,43	71,51	72,22	68,80	72,73
	Overlap	78,38	82,61	75,28	74,50	81,50
	ScaledColor	74,24	67,84	80,28	73,50	74,53
	ShapeTanimoto	78,33	74,31	81,81	78,54	78,79
	Tversky.d	73,62	73,44	74,58	71,42	76,69
	Tversky.q	74,95	74,52	76,94	73,25	77,30
RefC	ColorScore	72,86	71,00	74,72	72,21	72,94
	ComboScore	75,57	75,47	77,78	75,83	76,96

Table A.11: Cross-validated results of random forest

Crossval. Random Forest	15 Comp. Descriptors	A	A1	A0	Pr1	Pr0
	Overlap	79,10	81,50	75,00	76,34	80,83
	ScaledColor	74,95	69,05	79,58	75,65	73,22
	ShapeTanimoto	74,95	71,22	77,92	75,15	75,90
	Tversky.d	74,24	72,00	75,42	73,72	76,23
	Tversky.q	73,67	79,52	67,36	69,85	76,39
RefD	ColorScore	74,86	71,22	79,72	76,13	75,09
	ComboScore	80,33	80,05	80,83	77,88	81,00
	Overlap	77,71	80,39	73,89	75,09	78,57
	ScaledColor	76,33	70,34	83,19	79,51	75,63
	ShapeTanimoto	76,95	76,54	78,06	75,17	79,29
	Tversky.d	78,24	73,34	81,25	78,46	78,41
	Tversky.q	72,19	75,46	68,89	70,36	73,89
	vSURF	73,38	73,57	73,39	70,24	78,50
	vSURF	74,57	75,00	74,64	73,34	78,17
	vSURF	72,43	70,24	74,82	73,24	77,11
	vSURF	76,10	74,76	77,50	76,13	79,40
	pure vSURF	79,48	80,00	79,82	79,70	84,22
	VSA	74,19	72,14	75,71	74,55	75,38
	VSA	80,76	77,62	83,21	80,77	81,85
	VSA	72,05	67,86	76,07	73,69	73,25
	VSA	77,95	69,76	84,64	83,14	78,10
	pure VSA	80,10	75,00	84,64	81,71	80,36

Table A.11: Cross-validated results of random forest for study 2

Table A.12: Cross-validated results of second random forest for study 2

CV RF new	15 Comp Descriptors	A	A1	A0	Pr1	Pr0
RefA	ColorScore	73,48%	67,25%	79,86%	73,88%	72,53%
	ComboScore	74,34%	74,18%	76,67%	72,52%	76,56%
	Overlap	77,16%	80,98%	73,89%	73,15%	79,83%
	ScaledColor	72,81%	67,25%	78,75%	72,69%	72,77%
	ShapeTanimoto	76,29%	73,96%	79,44%	77,92%	77,31%
	Tversky.d	76,29%	75,94%	76,81%	75,49%	77,38%
	Tversky.q	77,52%	78,15%	76,39%	74,97%	79,22%
RefB	ColorScore	74,14%	68,48%	80,42%	73,07%	74,38%
	ComboScore	68,99%	67,84%	70,83%	66,79%	69,80%
	Overlap	76,95%	78,61%	76,94%	74,94%	79,13%
	ScaledColor	72,14%	65,30%	78,06%	74,22%	72,58%
	ShapeTanimoto	77,00%	71,11%	79,86%	78,21%	76,63%
	Tversky.d	77,92%	77,98%	76,25%	76,33%	80,64%
	Tversky.q	76,95%	80,02%	75,00%	74,60%	78,65%
RefC	ColorScore	74,90%	70,48%	78,47%	74,65%	73,83%
	ComboScore	74,86%	72,05%	79,03%	76,25%	75,21%
	Overlap	79,05%	84,83%	72,22%	75,14%	82,22%
	ScaledColor	75,52%	73,57%	78,61%	75,94%	75,30%
	ShapeTanimoto	77,67%	72,88%	81,39%	79,29%	77,94%
	Tversky.d	73,62%	72,00%	74,03%	72,23%	75,47%
	Tversky.q	72,29%	76,42%	68,06%	69,01%	74,15%
RefD	ColorScore	76,29%	71,22%	81,94%	78,44%	75,63%
	ComboScore	82,43%	82,05%	83,33%	81,40%	82,39%
	Overlap	79,10%	83,72%	73,89%	75,80%	80,97%
	ScaledColor	76,95%	70,11%	84,86%	80,94%	75,55%
	ShapeTanimoto	75,76%	75,58%	76,94%	74,28%	78,11%
	Tversky.d	77,74%	73,33%	81,25%	78,66%	77,47%
	Tversky.q	79,05%	81,15%	77,64%	76,96%	80,02%
RefA	vSURF	71,29%	69,05%	73,39%	69,09%	75,46%
RefB	vSURF	75,33%	75,24%	75,89%	76,26%	78,28%
RefC	vSURF	72,38%	70,24%	74,64%	69,05%	76,31%

Table A.12: Cross-validated results of second random forest for study 2

CV RF new	15 Comp Descriptors	A	A1	A0	Pr1	Pr0
RefD	vSURF	75,38%	73,10%	77,50%	75,65%	78,33%
	pure vSURF	80,14%	81,43%	79,82%	80,91%	85,74%
RefA	VSA	72,24%	69,05%	74,64%	72,57%	73,73%
RefB	VSA	73,48%	69,05%	77,32%	73,93%	74,68%
RefC	VSA	79,33%	69,76%	87,14%	85,39%	78,47%
RefD	VSA	74,95%	72,11%	77,54%	75,36%	77,62%
	pure VSA	81,68%	74,05%	88,57%	85,52%	79,98%
RefB	10ADME	73,62%	68,02%	78,92%	77,14%	72,17%
RefD	10ADME	74,95%	71,40%	81,14%	75,99%	75,39%
	pure 10ADME	74,29%	75,76%	72,53%	72,29%	78,60%

Table A.12: Cross-validated results of second random forest of study 2

Table A.13: Cross-validated results of tuned random forest of study 2

CV RF tuned	15 Comp Descriptors	A	A1	A0	Pr1	Pr0
RefA	ColorScore	0,73	0,67	0,80	0,74	0,73
	ComboScore	0,73	0,73	0,77	0,73	0,76
	Overlap	0,78	0,81	0,75	0,74	0,81
	ScaledColor	0,72	0,65	0,80	0,73	0,71
	ShapeTanimoto	0,78	0,75	0,81	0,80	0,78
	Tversky.d	0,76	0,74	0,78	0,76	0,76
	Tversky.q	0,76	0,77	0,77	0,74	0,78
RefB	ColorScore	0,74	0,68	0,80	0,73	0,74
	ComboScore	0,70	0,68	0,72	0,68	0,71
	Overlap	0,77	0,78	0,77	0,74	0,79
	ScaledColor	0,73	0,66	0,78	0,75	0,73
	ShapeTanimoto	0,78	0,69	0,84	0,80	0,77
	Tversky.d	0,77	0,76	0,76	0,76	0,80
	Tversky.q	0,76	0,81	0,74	0,74	0,79

Table A.13: Cross-validated results of tuned random forest of study 2

CV RF tuned	15 Comp Descriptors	A	A1	A0	Pr1	Pr0
RefC	ColorScore	0,75	0,70	0,78	0,75	0,74
	ComboScore	0,77	0,77	0,79	0,77	0,79
	Overlap	0,78	0,82	0,73	0,75	0,81
	ScaledColor	0,79	0,72	0,84	0,81	0,76
	ShapeTanimoto	0,76	0,72	0,80	0,77	0,77
	Tversky.d	0,76	0,73	0,78	0,77	0,78
	Tversky.q	0,74	0,75	0,72	0,72	0,75
RefD	ColorScore	0,76	0,70	0,83	0,79	0,75
	ComboScore	0,81	0,81	0,81	0,79	0,82
	Overlap	0,78	0,80	0,74	0,75	0,79
	ScaledColor	0,75	0,66	0,84	0,81	0,73
	ShapeTanimoto	0,79	0,77	0,82	0,80	0,80
	Tversky.d	0,78	0,70	0,85	0,83	0,76
	Tversky.q	0,78	0,79	0,78	0,77	0,78
RefA	vSURF	0,75	0,73	0,77	0,73	0,79
RefB	vSURF	0,74	0,72	0,76	0,74	0,77
RefC	vSURF	0,74	0,72	0,76	0,74	0,79
RefD	vSURF	0,76	0,75	0,78	0,76	0,79
	pure vSURF	0,80	0,81	0,80	0,81	0,86
RefA	VSA	0,74	0,70	0,77	0,75	0,75
RefB	VSA	0,72	0,69	0,76	0,74	0,74
RefC	VSA	0,79	0,70	0,86	0,84	0,78
RefD	VSA	0,81	0,78	0,83	0,80	0,82
	pure VSA	0,81	0,75	0,87	0,84	0,80
RefB	10ADME	0,74	0,72	0,77	0,75	0,73
RefD	10ADME	0,74	0,70	0,81	0,76	0,75
	pure 10ADME	0,74	0,77	0,71	0,72	0,80

Table A.13: Cross-validated results of tuned random forest of study 2

Table A.14: Cross-validated results of support vector machine of study 2

Crossval. Support Vector Machine	15 Comp. RBF Kernel Descriptors	A	A1	A0	Pr1	Pr0
RefA	ColorScore	74,76	74,08	75,97	73,86	76,16
	ComboScore	76,19	57,33	94,31	89,00	71,12
	Overlap	76,33	78,89	73,61	74,17	78,35
	ScaledColor	71,38	68,81	74,03	71,22	71,98
	ShapeTanimoto	76,86	73,27	79,44	78,71	75,20
	Tversky.d	76,90	78,60	75,56	72,93	80,77
	Tversky.q	75,48	75,48	72,78	72,41	75,53
RefB	ColorScore	72,90	60,71	85,69	77,81	70,80
	ComboScore	70,90	73,10	70,83	69,27	74,94
	Overlap	78,19	77,52	80,42	77,75	80,40
	ScaledColor	66,57	62,88	72,08	65,78	69,36
	ShapeTanimoto	77,05	71,03	84,72	81,95	75,82
	Tversky.d	74,24	78,92	69,31	71,08	75,99
	Tversky.q	78,29	78,94	75,97	76,65	78,63
RefC	ColorScore	73,38	75,39	70,69	69,75	76,77
	ComboScore	76,29	53,96	97,78	94,67	70,19
	Overlap	78,29	79,83	75,00	75,43	80,12
	ScaledColor	71,67	54,18	89,03	82,00	67,91
	ShapeTanimoto	78,33	70,88	87,22	84,67	76,09
	Tversky.d	79,71	77,02	83,61	81,83	79,10
	Tversky.q	77,57	72,44	81,25	76,94	77,14
RefD	ColorScore	76,24	64,79	89,86	85,64	73,69
	ComboScore	78,90	82,00	78,75	78,35	81,82
	Overlap	78,86	82,31	74,31	74,77	82,28
	ScaledColor	75,57	63,88	89,86	85,64	73,24
	ShapeTanimoto	77,57	76,07	79,86	78,52	77,43
	Tversky.d	78,81	79,71	76,81	76,04	80,14
	Tversky.q	76,90	80,96	71,94	72,22	80,12
	vSURF	77,48	74,29	79,82	76,83	79,32
	vSURF	76,67	73,81	79,46	77,59	80,07
	vSURF	76,52	72,86	79,64	77,21	79,42

Table A.14: Cross-validated results of support vector machine of study 2

Crossval. Support	15 Comp. Vector Machine RBF Kernel Descriptors	A	A1	A0	Pr1	Pr0
	vSURF	78,71	77,38	79,46	77,43	81,41
	pure vSURF	74,67	73,33	75,89	74,87	78,56
	VSA	84,90	83,81	86,07	84,96	86,53
	VSA	80,10	79,29	80,89	81,35	83,82
	VSA	83,62	83,81	83,57	82,33	86,37
	VSA	80,76	79,52	82,14	79,35	84,63
	pure VSA	81,52	80,71	82,32	81,45	84,69

Table A.14: Cross-validated results of support vector machine of study 2

Abstract

ABC (ATP-binding cassette) transporters represent membrane bound efflux pumps dependent on ATP with the most prominent member being ABCB1 or P-glycoprotein. This protein is placed at important junctions like the blood-brain barrier, the gut wall mucosa, the placenta and among others hepatobiliary pathways and consequently plays a major role in drug-drug interactions and multi-drug resistance. For this reason recognition of substrate properties and reliable labelling of non-substrates gain more and more importance.

In-house derived similarity based descriptors (SIBAR) were previously developed with focus on poly-specific proteins like ABCB1. These descriptors depend upon a set of reference compounds and consist of the calculated euclidian distance between descriptor values of the reference set and the training and test set. The number of final descriptors therefore is dependent on the number of compounds in the reference set. Four different reference sets have been derived and their usability is discussed. The first three reference sets are based on Tudor Oprea's chemography idea of satellite structures on the fringes of the chemical space. The fourth reference set is based on in-house results favouring a tailored reference set to the training set.

The focus of this work lay in the exploration of the 3D usability of the SIBAR descriptors and the impact of shape similarity based on a consistent data set. Also the establishment of a suitable reference set for a classification model for ABCB1 is discussed based on 240 highly diverse natural compounds.

In order to achieve this goal a variety of different descriptor types and machine learning approaches were performed. These include 2D descriptors, Labute's VSA descriptors, 3D Autocorrelation descriptors and VolSurf descriptors. To further explore the concept of shape similarity the parameters derived from the program ROCS of Openeye via shape overlay between two molecules were also used as descriptors. The machine learning approaches primarily encompass binary QSAR, support vector machine and random forest.

Results show that 2D descriptors compare very creditably with 3D or shape based methods and especially the VSA descriptors presented the best model so far with an overall accuracy of 83%. As appropriate reference set reference set B is preferred if a quick overview is necessary. For more detailed analysis of one's data a more time-consuming tailored reference set is the reference set of choice. The overall results were satisfying and especially some of the ROCS parameters showed good performance.

Zusammenfassung

ABC (ATP-binding cassette) Transporter stellen membrangebundene Ausscheidepumpen dar mit dem prominentesten Mitglied ABCB1 oder P-glycoprotein. Dieses Protein ist an wichtigen Stellen wie der Blut-Hirn-Schranke, der Magenwandschleimhaut, der Plazenta und anderen Leber-Stoffwechselprozessen lokalisiert und spielt daher eine große Rolle bei Medikamentenwechselwirkungen und auch bei Mehrfachresistenzen von verschiedenen Arzneimittel. Deshalb ist die Erkennung von Substrat-Eigenschaften und eine zuverlässige Erkennung von Nichtsubstraten so bedeutsam.

In unserer Gruppe wurden ähnlichkeitsbasierte Deskriptoren (SIBAR) mit Fokus auf solch polyspezifische Proteine wie ABCB1 entwickelt. Diese Deskriptoren sind abhängig von einem Referenzsatz und bestehen aus den berechneten Euklidischen Distanzen in Bezug auf den Referenzsatz und die entsprechenden Training- und Testsätze. Ihre Anzahl ist daher abhängig von der Zahl der Moleküle im Referenzsatz. Vier verschiedene Referenzsätze wurden abgeleitet und ihre Anwendbarkeit diskutiert. Die ersten drei Referenzsätze basieren auf Tudor Oprea's Idee von Chemographie in Form von Satelliten-Strukturen am Rande des chemischen Raums. Der vierte Referenzsatz wurde entsprechend vorherigen Ergebnissen in unserer Gruppe auf den spezifischen Trainingssatz zugeschnitten.

Der Fokus dieser Arbeit liegt vor allem in der Erforschung der Anwendbarkeit von 3D Deskriptoren in Bezug auf SIBAR und des Nutzen von Form-Ähnlichkeit basierend auf einem konsistenten Datensatz. Außerdem wird die Entwicklung eines geeigneten Referenzsatzes zur Erstellenung eines Klassifikationsmodells für ABCB1 basierend auf 240 sehr diversen Naturprodukten diskutiert.

Um dieses Ziel zu erreichen wurde eine Vielzahl von verschiedenen Deskriptortypen und auch verschiedene Klassifikationsansätze verwendet. Diese beinhalten 2D-, VSA-, 3D Autokorrelations- und VolSurf- Deskriptoren. Um außerdem noch das Konzept von Form-Ähnlichkeit zu untersuchen wurden die Parameter, die durch das Programm ROCS von Openeye bei der Übereinanderlegung von zwei Molekülen errechnet worden waren, als Deskriptoren verwendet. Die verwendeten Klassifikationsmodelle beinhalten binary QSAR, Support Vector Maschinen und random forest.

Die Ergebnisse zeigen, dass 2D Deskriptoren im Vergleich zu 3D oder Form-basierten Methoden sehr gut abschneiden und besonders mit VSA Deskriptoren konnte das beste Modell mit 83% Genauigkeit erreicht werden. Abschließend bemerkt ist Referenzsatz B sehr gut geeignet um einen schnellen Überblick über den Datensatz zu ermöglichen, aber bei detaillierter Betrachtung ist ein spezifisch zugeschnittener Referenzsatz zu bevorzugen. In

ihrer Gesamtheit waren die Ergebnisse zufriedenstellend und besonders einige von ROCS errechneten Parameter zeigten gute Leistung.

Curriculum Vitae

Personal data

Rita Schwaha
Date of birth: 23.08.1979
Languages: German, English, French, Italian

Education

since 03/06	Doctoral studies of natural sciences University of Vienna, Austria. Thesis:"Similarity based classification studies for prediction of ABCB1 (P-glycoprotein) substrates and non-substrates" supervised by Univ.-Prof. Dr. Gerhard F. Ecker.
03/08 - 06/10	Postgraduate course of chinese diagnostics and pharmatherapy
07/04 - 06/05	Pre-registration year for pharmacists at the combined hospital and community pharmacy of the Barmherzigen Brüder, Linz, Austria
10/1998 - 04/04	Studies of pharmaceutcial sciences University of Innsbruck, Austria. Diploma Thesis "Sesquiterpenlactones in *Cicerbita alpina* and *Leontodon rigens*" supervised by Dr. Christian Zidorn and Univ.-Prof. Dr. Hermann Stuppner.
1998	Advanced study course in italian language at the university of Bergamo, Italy
1998	A levels passed summa cum laude
1990 - 1998	Grammar School in Linz, Austria
1985-1989	Primary School in Linz, Austria

Professional life

since 07/2006	Pharmacist at the Prinz Eugen pharmacy in Linz, Austria
01/07 - 12/09	Scientific assistant in the Pharmacoinformatics Research group of Univ.-Prof. Dr. Gerhard F. Ecker, University of Vienna, Austria financed by the FWF (Austrian scientific fund)

10/2005 - 07/06	Scientific assistant at Austrian Research Center Seibersdorf, Austria in the field of Molecular Modelling
07/05 - 10/05	Registered pharmacist at the combined hospital and community pharmacy of the Barmherzigen Brüder, Linz, Austria.
07/04 - 06/05	Pre-registration year at the combined hospital and community pharmacy of the Barmherzigen Brüder, Linz, Austria.

Internships

1996	four weeks at the Voest Alpine Industrieanlagenbau in Linz, Austria
1997	four weeks at the Voest Alpine Industrieanlagenbau in Linz, Austria
1998	four weeks at the VAI Impianti in Bergamo, Italy
1999	four weeks at the VAI Impianti in Bergamo, Italy
2000	four weeks at the hospital pharmacy of the community hospital in Linz, Austria
2001	two months at the vitamins research laboratory of Hoffmann La Roche in Basel, Switzerland
2002	four weeks at the quality management department of DSM Fine Chemicals in Linz, Austria
2003	four weeks at the community pharmacy of Lloydspharmacies in Coventry, Great Britain

Publications

Publications in journals

C. Zidorn, E.P. Ellmerer, G. Konwalinka, K.H. Ongania, R. Schwaha, R. Greil, K. Joehrer, N.B. Perry, H. Stuppner. 1,10-Epoxyhypocretenolides from the Azorean Endemic *Leontodon rigens* (Asteraceae). Letters in Organic Chemistry, 2(5): 461-464 (2005).

C. Zidorn, R. Schwaha, E.P. Ellmerer, H. Stuppner. On the occurrence of sonchuside A in *Cicerbita alpina* and its chemosystematic significance. Journal of the Serbian Chemical Society, 70(2): 171 - 175 (2005).

R. Schwaha, G.F. Ecker. The Similarity Principle - New Trends and Applications in Ligand-Based Drug Discovery and ADMET Profiling, Scientia Pharmaceutica, 76: 5-18 (2008).

M.A. Demel, R. Schwaha, O. Krämer, P. Ettmayer, E. Haaksma, G.F. Ecker. *In silico* prediction of substrate properties for ABC multidrug transporters. Expert Opinion on Drug Metabolism and Toxicology, 4(9): 1167-1180 (2008).

R. Schwaha, G.F. Ecker. Similarity Based Descriptors - Useful for classification of Substrates of the Human Multidrug Transporter P-Glycoprotein? Journal of QSAR and Computational Sciences 28 (8), 835-839 (2009).

R. Schwaha, G.F. Ecker. Use of shape similarities for the classification of P-glycoprotein substrates and nonsubstrates. Future Medicinal Chemistry, 3(9): 1117-28 (2011).

Talks

R. Schwaha, G.F. Ecker: SIBAR descriptors and Support Vector Machine for ABCB1 substrate prediction. Presented at the 21. Scientific congress of the Austrian Pharmaceutical Society, April 16-18th 2009.

Conference contributions

T. Stockner, R. Schwaha, M. Jakusch. Model of nuclear receptor ligand binding; ISQBP-Presidents Meeting, Strasbourg, June 24-27th, 2006. - presented by Dr. Thomas Stockner

R. Schwaha, M. Demel, D. Kaiser, G.F. Ecker. Classification of P-Glycoprotein-substrates using similarity based descriptors; Poster at the annual convention of the German Pharmaceutical Society in Erlangen, Germany, October 10-13th 2007

G. F. Ecker, R. Schwaha, M. Demel, P. Chiba. Pharmacoinformatic approaches to Target P-glycoprotein: From inhibitor design to substrate prediction; Poster at the 235th ACS Meeting in New Orleans, United States, April 6-10th 2008 – presented by Univ.-Prof. Dr. Gerhard F. Ecker

A.G.K. Janecek, R. Schwaha, M.A. Demel, W.N. Gansterer, G.F. Ecker. Applying Web Search Algorithms for ADME/TOX Profiling of ABCB1 (P-GP) Substrates; Poster at the XXth International Meeting of Medicinal Chemistry in Vienna, Austria, August 31st - September 4th 2008. - presented by DI Andreas Janecek

R. Schwaha, G. F. Ecker. Classification of ABCB1 Substrates using Similarity Based Descriptors; Poster at the XXth International Meeting of Medicinal Chemistry in Vienna, Österreich, August 31st - September 4th 2008.

R. Schwaha, G.F. Ecker. Similarity Based Descriptors for classification of ABCB1 substrates. Poster at the EuroQSAR 2008 in Uppsala, Schweden, 21 - 26th September 2008.

R. Schwaha, G.F. Ecker. 2D and 3D-Similarity based classification systems for substrate prediction of ABCB1 at the National Meeting of the American Chemical Society in Washington DC, USA, August 16 - 20th, 2009.

Vienna, May 22, 2013