universität wien

DISSERTATION

Titel der Dissertation

„In-silico Models for the Characterization of Compounds interfering with clinical relevant ABC-Multidrug-Transporters“

verfasst von

Mag.pharm. Michael Alexander Demel

angestrebter akademischer Grad

Doktor der Naturwissenschaften (Dr.rer.nat.)

Wien, 2013

Studienkennzahl lt. Studienblatt:	A 796 610 449
Dissertationsgebiet lt. Studienblatt:	Pharmazie
Betreut von:	Univ.-Prof. Mag. Dr. Gerhard F. Ecker

"ABC-Transporters: Friend or Foe?"

Acrylic on linen

PYP2.0 – Paint Your PhD contest,

M.A. Demel,

Vienna, May 2012

"I can`t understand why people are frightened of new ideas.

I`m frightened of the old ones."

John Cage (1912-1992)

"Prediction is very difficult, especially about the future."

attributed most often to Nils Bohr, but also to
Mark Twain, Albert Einstein, or Yogi Berra

Publications enclosed in this thesis:

Chapter 9:

Demel M., Schwaha R., Krämer O., Ettmayer P., Haaksma E., Ecker G.F.

In Silico Prediction of Substrate Properties for ABC-Multidrug-Transporters.

Expert Opinion on Drug Metabolism and Toxicology, 2008, 4(9):1167-1180.

Chapter 10:

Demel M., Krämer O., Ettmayer P., Haaksma E., Ecker G.F.

Predicting Ligand-Interaction with ABC-Transporters in ADME.

Chemistry and Biodiversity, 2009, 6, 1960-1969.

Chapter 11:

Demel M., Janecek A.G.K., Gansterer W.N., Ecker G.F.

Comparison of Contemporary Feature Selection Algorithms: Application to the Classification of ABC-Transporter Substrates.

QSAR and Combinatorial Sciences, 2009, 28, 1087-1091.

Chapter 12:

Demel M., Krämer O., Ettmayer P., Haaksma E., Ecker G.F.

Ensemble Rule-Based Classification of Substrates of the Human ABC-Transporter ABCB1 Using Simple Physicochemical Descriptors.

Molecular Informatics, 2010, 29, 233-242.

Other publications in the same field but not enclosed in this thesis:

Demel M., Janecek A.G.K., Thai K.-M., Gansterer W.N., Ecker G.F.

Predictive QSAR Models for Polypecific Drug Targets: The Importance of Feature Selection.

Current Computer-aided Drug Design, 2008, 4:91-110.

Janecek A.G.K., Gansterer W.N., **Demel M.**, Ecker G.F.

On the Relationship between Feature Selection and Classification Accuracy.

Journal of Machine Learning Research: Workshop and Conference Proceedings, 2008, 4, 90-105.

Summary of Contributions to the Field

Human ATP-binding-cassette-(ABC)-transporters - especially ABCB1 (P-glycoprotein, P-gp, MDR1), the best characterized drug transporter of this protein family - act as cellular multidrug exporters for a variety of structurally and pharmacologically unrelated drugs. Their hallmark feature is to exhibit a rather polyspecific ligand recognition pattern and to extrude their substrates out of living cells against a steep concentration gradient. Furthermore, they are well known for representing important determinants for drug efficacy and drug safety. More than three decades of extensive research has identified many clinically used drugs as substrates of ABC-transporters. Substrates of these multi-drug efflux pumps tend to show reduced clinical effectiveness on the one hand but can also give rise to serious drug-drug interactions on the other hand when exposed to tissues that show high expression levels of ABC-transporters. ABC-transporter overexpression in cancer cells is one of the main reasons for diminished treatment response towards many different cytostatic treatment regimens in cancer therapy, because of reduced intracellular drug concentrations. This phenomenon is also known as multidrug resistance (MDR) and is associated with poor survival rates. However, cancer cells are not the only cells showing high levels of ABC-transporters. Tissues that determine the absorption, distribution and elimination of drugs are also enriched in these efflux pumps. Therefore, the endogeneous presence of these drug exporters is key for the pharmacokinetic fate of substrate molecules. Summarising the critical role of ABC-transporters in cancer cells and their important contribution to the ADMET profile of various drugs, makes clear that the correct identification of substrates of these transporters is of high medical significance and a critical prerequisite for the successful clinical development of novel drugs.

The fact that research and development of novel drugs is commonly associated with poor success rates and tremendous costs raises the importance for acknowledging efficacy and safety issues as early as possible in a drug development project. *In-silico* methods are highly recognized as important and competent means to facilitate and guide decision making early in the drug discovery process and furthermore harbour the potential to substantially reduce drug development costs. *In-silico* classification systems are also nowadays routinely applied in early ADMET profiling of drug candidates. Furthermore, they were also applied to model ABCB1 substrates and non-substrates by employing QSAR studies, docking into homology models, pharmacophore modelling, but also machine learning approaches. QSAR studies usually rely upon the presence of a congeneric series of compounds. Considering the broad substrate diversity of ABCB1 and other relevant transporters, limits the application of QSAR studies for this modelling task. Although docking is a successful tool for virtual screening, it also shows some limitations with respect to categorize ligands of ABCB1. First of all, the target structure can currently only be estimated by a homology model, since a high-resolution crystal structure of human ABCB1 has not been published to date. Second, the presence of a huge binding cavity in ABCB1 that is constituted of multiple binding sites, renders docking into an ABCB1 model a rather difficult and labour intensive approach. Ligand-based pharmacophore models and machine learning approaches constitute currently the state-of-the-art methods to model ABCB1 substrate liability. However, most of the models published so

far are only based on rather small compound sets and many of the machine learning models are based on highly complex algorithms that can only be interpreted to a limited extend.

In this thesis different *in-silico* ligand-based classification models, which are methodologically based on machine learning theory, for ligands of ABC-transporters are presented and critically appraised. Many different machine learning algorithms are nowadays available. However, these methods differ remarkably with respect to their predictive performance and their algorithmic complexity. Traditionally, increasing algorithmic complexity positively influences model performance but at the same time also severely exacerbates model interpretation and retrieval of useful information, which often compromises the identification of underlying chemical concepts that could provide useful guidance for the medicinal chemist in prospective research projects. Conversely, simple machine learning methods often provide tools for model interpretation, but usually tend to lack predictive performance that would assure the necessary robustness and generalizability of the model.

It is the primary objective of this work to establish classification systems that combine good and reliable classification performance but also allow for meaningful model interpretation to enable the characterization of physico-chemical and structural properties of ligands of ABC-transporters derived from different sources. From a methodological viewpoint, different data pre-processing methods (mainly derived from the field of feature selection), different modelling algorithms as well as sophisticated data post-processing methods are applied and compared to accomplish this primary objective. Data pre-processing in the context of feature selection aims to remove redundant information form the data set and to reduce model complexity by reducing the dimensionality of the data. Contemporary machine learning algorithms, belonging to the class of ensemble methods, appear to be versatile tools for achieving the primary objective of this thesis, since they aim to unify classification performance and interpretability. Furthermore, data post-processing often utilizes visualization techniques which help to summarize model information in a compact way and can also serve for assessing the prediction reliability of the model when applied to real-life problems. In this thesis data sets from different ressources are used for modelling purposes. Some are derived from the public domain, whereas others are proprietary molecules that resemble a comprehensive collection of oncological hit and lead molecules. The feasibility of the different data sets is also discussed.

Most of the models presented in this thesis address the problem to reliably discriminate between ABCB1 substrate and non-substrates. However, initial attempts to provide classification models for two other clinically relevant ABC-transporters, namely ABCC1 and ABCG2, are also presented. Additionally, the novel promising class of MDR-selective molecules (cytotoxic compounds that kill ABC-transporter overexpressing cancer cells more efficiently than transporter-naïve cancer cells), that might serve as innovative starting point for the development of future anti-cancer agents, are also characterized by means of machine learning algorithms but also by using network-like graphs.

The table below gives a short overview on the different modelling approaches anticipated in this thesis.

Quantitative Summary of the individual modelling efforts conducted in this thesis:

chapter	primary objective	modelling problem	data set	methods	type of validation
Chapter 10	**post-processing:** interpretability of a classification model	ABCB1 substrates/ non-substrates	molecules with binnarized ABCB1-liability data compiled from literature sources and the NCI60 drug sensitivity screen	**ensemble modelling method:** RuleFit algorithm	**internal validation** (10-fold cross-validation)
Chapter 11	**pre-processing:** comparison of different feature selection methods	ABCB1 ABCC1 ABCG2 substrates/ non-substrates	three data set from the NCI60 drug sensitivity screen comprising binary activity data for ABCB1, ABCC1, ABCG2 substrates/non-substrates	**feature selection algorithms:** filters incl. feature ranking, wrappers **modelling alogorithm:** *k*NN	**internal validation** (10-fold cross-validation)
Chapter 12	**modelling:** application and appraisal of the RuleFit algorithm **interpretability:** ABCB1-substrate classification rule	ABCB1 substrates/non-substrates	comprehensive proprietary data set with binnarized activity measured in a cytotoxicity assay	**ensemble modelling method:** RuleFit algorithm	**external validation** using proprietary molecules
Chapter 13	**post-processing**: interpretability and applicability domain estimation	ABCB1 substrates/non-substrates	proprietary data literature data, NCI60 drug sensitivity data	**ensemble modelling method:** Random Forests	**external validation** using proprietary molecules
Chapter 14	**pre-processing:** descriptor development	ABCB1 substrates/non-substrates	public available data set compiled from different literature sources	**descriptors:** MoSS-guided FP-SIBAR **ensemble modelling method:** Random Forests	**internal validation** (10-fold cross-validation)
Chapter 15	**modelling:** classification of MDR-selective molecules using three different classifiers	MDR-selective compounds	86 molecules with binnarized selectivity values	Random Forests, Boosting, Support Vector machine	**external validation:** pre-defined test set
Chapter 16	**interpretation:** exploration of selectivity determinants within a series of thiosemicarbazones selectively killing MDR cells	MDR-selective compounds	41 thiosemicarbazones with continuous IC_{50} values and selectivity ratios	structural exploration using Network-like similarity graphs	NA*

*NA = not applicable

In Chapter 10 the development of an Ensemble-Rule based model for classifying a public available data set of ABCB1 substrates and non-substrates is presented. The applied algorithm is Friedman`s RuleFit algorithm, a novel method that was first described in 2004 and concentrates on model interpretation. The model is validated using 10-fold crossvalidation and the interpretation of this model highlights the role of simple physico-chemical descriptors for characterising ABCB1 substrates. The same algorithm is also applied to a much larger and more diverse data set of proprietary compounds from a comprehensive hit/lead library (Chapter 12). The models in Chapter 12 were also validated on an external test set, which consists of more than 1000 molecules. It is shown that models based on large, diverse training sets performed best and that substrates tend to show a higher number of hydrogen-bond acceptors and are more flexible. Furthermore, the importance of lipophilicity

is also reflected by this model. RuleFit proves to be a well-suited tool for modelling ABC-transporter substrates.

As outlined above, another way to facilitate model interpretation is to apply sophisticated feature selection algorithms as pre-processing tools prior to the actual modelling process. In Chapter 11 five contemporary feature selection algorithms are compared with respect to their predictive performance, but also with respect to the suitability of the retrieved feature subsets for interpretation purposes. Data sets for ABCB1, ABCC1 and ABCG2 are used for this benchmarking study. The results highlight that feature selection methods, that incorporate the modelling algorithm into the selection process – so-called wrappers – return the best results. The interpretation of the best performing feature subsets suggests that descriptors based on van-der-Waals surface area projections provide in general good tools for modelling the substrates of all the three transporters under consideration.

Chapters 13, 14 and 15 utilize Breiman`s Random Forests algorithm. A widely used tool that also belongs to the class of ensemble learners and provides integrated functions to analyse complex models and non-linear data relationships. In Chapter 14 a different data pre-processing strategy is anticipated. Instead of trying to select an optimal subset of calculated descriptors (Chapter 11), the focus here is to develop a subset of new descriptors. These new descriptors, are based on similarity calculations to a set of reference molecules, and represent an extension of the previously reported SIBAR approach. The driving force for the development and evaluation of these new descriptors is to generate a set of features that encode chemical concepts in an intuitive way to improve model interpretation and to facilitiate extraction of useful SAR information. It is shown that the extended SIBAR descriptors exhibit performances comparable to structural fingerprints on modelling a set of public available ABCB1 substrates/non-substrates. The ease to interpret these novel descriptors is demonstrated on an example. Summarising these experiments, it must be concluded that these novel SIBAR descriptors are much more complex to generate than conventional descriptors. The wide applicability of this new approach shall be further evaluated on different data sets.

Chapter 13 focusses on the external validity of ABCB1 substrate/non-substrate classification models by utilising different variants of carefully compiled training sets. Some of the training sets constitute marketed drugs, another one is mainly composed of natural products, whereas the last resembles a diverse collection of proprietary inherent cytotoxic molecules. In total 27 different Random Forests models are presented in Chapter 13 and are used to predict an external test set. It is shown that some models, despite encouraging training performance, clearly fail to correctly classify the test compounds. This unravels a potential "efficacy-effectiveness" gap; i.e. models fail when applied to real-life situations. In an attempt to overcome this "efficacy-effectiveness"-gap, two different measures for assessing the applicability domain in the context of model post-processing are employed. The comparison of these two measures suggests that a distance-based measure performs better than the simpler range-based measure. However, the application of these post-processing techniques cannot fully explain why certain models fail to predict the validation set. Considering, that the different data sets origin from different sources and are sometimes based on different

surrogate endpoints that only approximate ABC-transporter mediated transport, suggests a different explanation for the poor performance of some models: predictive models do not only fail, because test compounds are out of chemical space, but might also fail because test compounds are out of biological/pharmacological space. In other words, the combination of different (surrogate) assay data from different experimental designs, that eventually might capture different aspects of ABC-transporter pharmacology, might be a reasonable explanation for the failure of certain models.

Chapter 15 and Chapter 16 are concentrating on a different pharmacological class than ABC-transporter substrates. Here, the focus is to characterize molecules that display higher cytotoxicity values when being exposed to ABC-tranporter expressing cells than to ABC-transporter-naïve cells. These MDR-selective molecules are characterized by means of classification models in Chapter 15. Therein, three different machine learning algorithms are compared. It is shown that Random Forests performs better than the other two methods on an external test set. The best model is interpreted by means of partial dependence plots. In Chapter 16, a congeneric series of MDR-selective isatin-β-thiosemicarbazones is analysed retrospectively by employing network-like similarity graphs to explore the molecular basis of the selectivity of this class towards ABCB1-overexpressing cells.

In summary, the individual chapters contained herein cover examples of the application of different machine learning models for compounds interfering with clinical relevant ABC-transporters.

Structure of this Thesis

This thesis is written in accordance with the guideline "Kumulative Dissertation" of the University of Vienna. It is structured into five parts. Part A introduces the reader into the physiological and pathological role of human ABC-transporters, thereby setting the basis for the understanding of the contribution of these efflux pumps to biomedicine and contemporary pharmacology. Part B provides the theoretical framework for the computational methods applied in this thesis. It categorizes the different types of pharmacoinformatic approaches that are currently applied in the field of drug discovery and design and also explains the fundamentals of the applied machine learning algorithms as well as methods of their validation. Part A and Part B close with a comprehensive reference section. Part C defines the particular objectives of this thesis.

Part D constitutes the result section of this work. It contains four published articles. The first two articles review the recent progress in the field. The first article focusses on the different pharmacological assays that are employed to measure transport by ABC-transporters. The output of these assays is of particular importance, since it is the basis for the response variable used in the development of machine learning models. This article gives also an overview of previously published *in-silico* models for ABC-transporter ligands. The second article focusses on ligand-based, simple rules that were previously established to characterize efflux pump ligands. This article also features a classification rule model for a comprehensive set of public available ABC-transporter substrates and non-substrates. Addionally, the available crystal structures of ABC-transporters are discussed. The next two chapters comprise peer-reviewed original articles. The third article concentrates on the application of sophisticated data pre-processing methods and compares and discusses five different feature selection algorithms. The fourth chapter of Part D outlines an ensemble rule-based model for a set of more than 1800 proprietary drug-like molecules. Here the focus is on the interpretation of this ensemble model. Additionally, Part D contains four manuscripts that contain further results of this thesis. These manuscripts are not considered for journal publication. The first manuscript focusses on the capability of a Random Forests model to predict external test compounds retrieved from different ressources and discusses approaches to estimate the applicability domain of a machine learning classification model. The next chapter concentrates on descriptor development. A novel approach to generate SIBAR descriptors is presented and critically appraised with respect to the classification performance of these extended SIBAR descriptors but also with respect to their suitability to interpret complex machine learning models. The last two chapters of Part D describe preliminary efforts to model the recently characterized and novel pharmacological class of MDR-selective molecules, i.e. compounds that exhibit higher levels of cytotoxicity to cancer cells in presence of ABC-transporters. Each chapter in the result section contains a declaration on the individual contributions of all applicable authors to the respective publication or manuscript. Part E summarizes and discusses the different modelling efforts presented in this thesis.

The appendix at the end of this thesis gives an overview on the poster conference contributions at European Conferences in the field of Medicinal Chemistry that were presented in the course of this thesis. Finally, the Conflict of Interest Statement, the Curriculum Vitae, as well as a abstract (English, German) is appended.

Table of Content

PART A: PHARMACOLOGICAL BACKGROUND

Chapter 1: Historical Overview on ABC-Transporter Research: From Target to Anti-target - and back?

ABC-transporters represent membrane-bound transport proteins with multi-facetted physiological, pathological but also pharmacological roles. The human genome encodes 48 ABC-transporter genes that are categorized on basis of their protein sequence similarity into 7 subfamilies. Human multidrug ABC-transporters embody highly complex molecular machines that extrude a variety of chemicals, drugs and xenobiotics out of cells against a concentration gradient. Many of these have been shown to play a significant role in drug safety and efficacy. Additionally, at least fifteen of these have been associated with clinical features of multidrug resistance (MDR) in oncology. As a result there is now an enormous large body of research results that describes and analyses the interaction of various drugs and clinical candidates with mammalian multidrug efflux pumps. The flowchart below highlights the main historical discoveries in the field of ABC-transporter research.

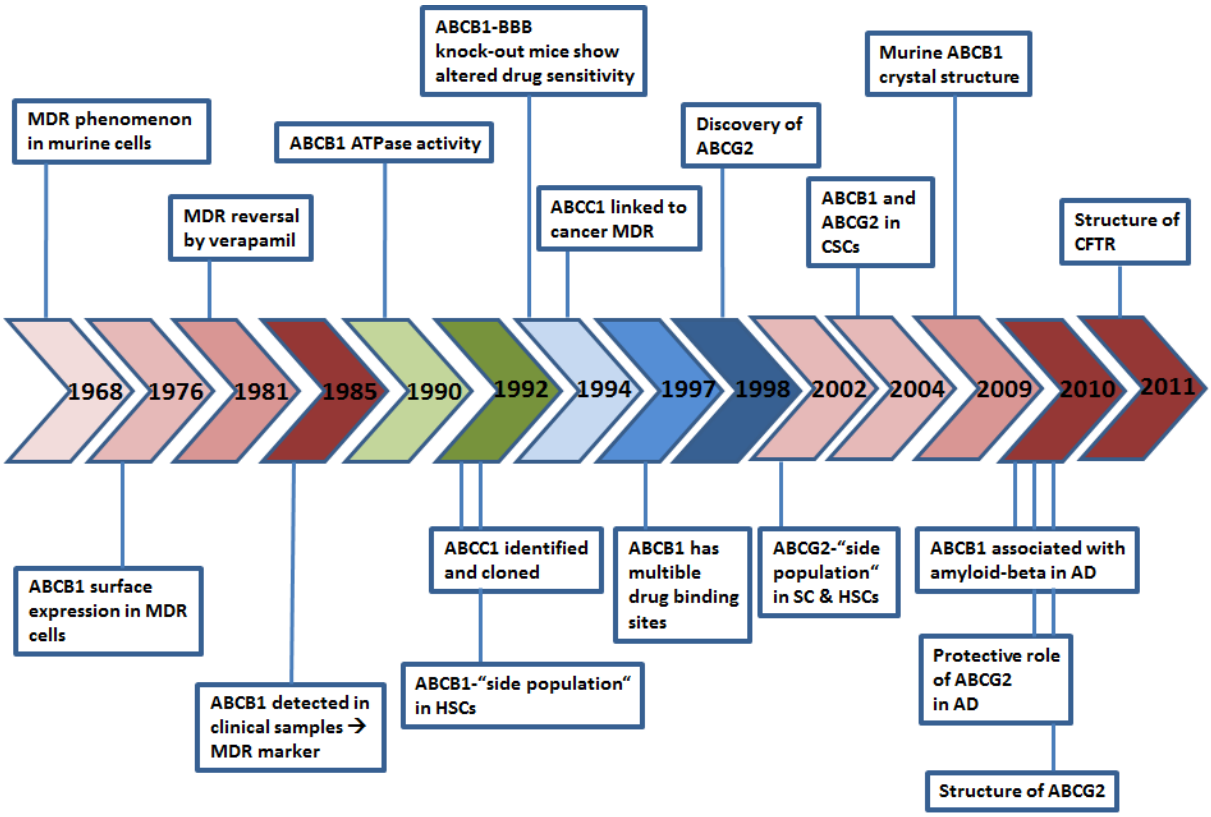


Figure 1: Landmark discoveries in the ABC-transporter field. Selected findings that outline the role of ABC-transporters in pathology and pharmacology (1–19). MDR = multi-drug resistant, BBB = bllod-brain barrier, HSC = hematopoietic stem cell, SC = stem cell, CSC = cancer stem cell, AD = Alzheimer`s disease.

From this figure it becomes apparent that ABC-transporters have now been studied for many decades. The most extensively characterized members are ABCB1 (P-glycoprotein, P-gp, MDR1), ABCC1 (MRP1) and ABCG2 (also known as BCRP or MXR). All of them are associated with MDR in many different types of malignancies, such as prostate, lung and breast cancer. ABCB1 and ABCC1 transport overlapping, mainly hydrophobic, but distinct classes of molecules. ABCG2 was also shown to transport a wide range of amphipathic substrates, including anionic species. Hence, all these efflux pumps are notorious for their promiscuous substrate recognition pattern. The initially recognized elevated expression of

these transporters, which results in reduced drug accumulation and consecutive decreased drug effectiveness in cancer cells, rendered them attractive *targets* for oncology. It has now been more than three decades since the discovery of ABCB1 and many inhibitors that were specifically designed with the objective to reverse this MDR phenotype, have been evaluated in clinical trials. However, so far the results of these trials were disappointing and the concept of direct transporter inhibition could not be translated into clinical practice (20). In the early 90`s of the last century, the role of ABC-transporters at tissue barriers such as the blood-brain barrier, the intestine, the hepatocytes and cells of the kidney was increasingly recognized and acknowledged. The realization that these proteins protect the organism against xenobiotics, but also interfere with the passage of drugs through tissues and hence can critically influence the pharmacokinetic profile of drugs, has led to the notion that ABC-transporters are indeed *anti-targets*. Consequently, it became one of the main research endeavors to deliberately design-out ABC-transporter substrate properties in order to optimize the ADMET profile of drug candidates. However, in the last century additional roles of ABC-transporters have been discovered. For instance, ABC-transporters have been associated with the onset of several neurodegenerative diseases and also seem to play putative crucial roles in inherent MDR cancer (stem) cells (21,22). Although these phenomena are far from being completely understood, they pave the way for the development of new strategies to improve future therapies in these indication areas. One such new promising strategy is nowadays to exploit the MDR phenotype of cancer cells by identifying and designing drugs that *selectively target* MDR cells over the non-resistant parental cells (23).

The following section gives a brief overview on ABC-transporters, their physiological roles, as well as their contribution to various diseases.

Chapter 2: ABC-Transporters in Health & Disease

Role of ABC-Transporters in Health

Intensive biochemical investigations have highlighted multiple physiological roles of several ABC-transporters encoded in the human genome. ABC-transporters are expressed under physiological conditions at tissue barriers and organs responsible for drug absorbtion, distribution or elimination (24). Hence, they constitute an effective "chemo-defence"-system that can be regarded (as suggested by B. Sarkadi) as a "third arm" of the human immune system (25–27). ABCB1 and ABCG2 are both expressed at the apical side of intestinal cells, and are oriented in that way, that they export substrates out of epithelial cells back into the lumen, thereby reducing further penetration and limiting oral bioavailability of various drugs. Recently, it was additionally shown that ABCC2 is also expressed at the same site. ABCG2 is additionally expressed at the blood-testis barrier and the maternal-fetal barrier (28). It also extrudes porphyrins from haematopoeitic cells and hepatocytes, and is responsible for the secreting riboflavin and vitamin K into breast milk (29). Additionally, ABCB1 alongside with ABCC2 and ABCC4 are also expressed in the tubular cells of the kidney at the luminal side. There, these exporters facilitate the secretion of drugs into the urine. In hepatocytes, ABCB1, ABCG2, and ABCB11 are expressed at the canalicular or apical side whereas the other ABC-transporters ABCC2, ABCC4, and ABCC6 are localized at the basolateral hepatocytic membrane. ABCG5 and ABCG8 are also present in the canalicular lumen and are responsible for transport of plant sterols (30). Furthermore, ABCB1 is expressed in the placenta in a stage-specific manner (31). It is important to note that these organs also express several uptake transporters of the organic anion transporter superfamily.

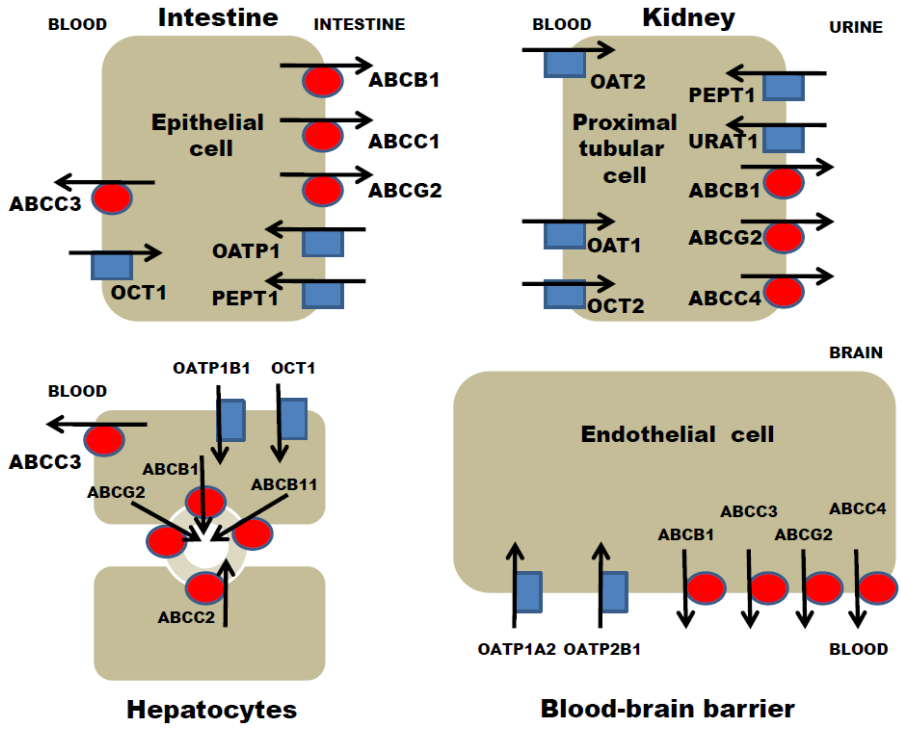


Figure 2: Schematic localization of transporters relevant for the disposition of various drugs. Efflux-pumps of the ABC-drug transporter family are localized at physiological tissues that represent barriers for drug transport. Members of the

ABC-transporter family are shown in red circles. Other transporters from other families are shown in blue squares (see text below). Modified and adapted from (24).

From Figure 2 it becomes apparent that ABC-transporters play important roles in the **A**bsorption, the **D**istribution and the **E**limination phases of the **ADME** paradigm. For **M**etabolisation enzymes belonging to the cytochrome P450 family are primarily responsiple.

The role of ABC-transporters in influencing drug disposition has been confirmed in knock-out mice that had remarkable defects in drug distribution (32). The potential of these transporters to influence the ADME-profile of drugs is of utmost interest to contemporary pharmacology. However, ABCB1 and other transporters are also localized at other cells which suggests that there they exhibit additional physiological functions. For instance, ABCB1 is also expressed on cells of the adrenal gland, hematopoietic stem cells, natural killer cells, dentritic cells and CD8+ T lymphocytes (33,34). Especially, immune cells show a high expression level of ABCB1 (but also of other members of this subfamily such as ABCB4/TAP). Several studies have shown that inhibition of ABCB1 or *silencing* via antisense oligonucleotides confers a reduction regarding the activity of these immune cells (35,36). Furthermore, several investigators have provided profound evidence that ABC-transporters are involved in the transport of several cytokines out of activated lymphocytes (37,38). Another report has pointed out that ABCB1 contributes to dentritic cell migration by transporting IL-1β and TNFα (34).

Additional roles of ABCB1 have also been discovered in recent years. It has been proposed that ABCB1 plays a fundamental role in regulating apoptosis (39,40). Robinson et al. showed that cells transfected with ABCB1 were resistant to apoptosis after serum starvation and this could be reversed after addition of verapamil. This indicates that the transport function of ABCB1 protects cells from death induced by growth-factor withdrawal (41). Biochemical analyses have revealed that ABCB1 prevents caspase mediated apoptosis; a result of UV radiation or Fas ligation (42,43). Functional ABCB1 is able to interfere with the proteolytic activation of caspase 3 and caspase 8 (44). This effect was reversed by anti-ABCB1 monoclonal antibodies. It was concluded that ABCB1 protects against caspase-dependent, but not caspase-independent cell death (45).

Other ABC-transporter members have also been shown to confer physiological functions. Other members of the ABCB-family, like ABCB4 and ABCB11 are involved in bile formation and are also key for the transport of bile acids (46,47). Members of the ABCA-family are crucial for maintaining lipid homeostasis. ABCA1 is responsible for the generation of HDL and ABCA2 is responsible for generating myelin sheats (48–52). Especially, ABCA1 is a key molecule for cholesterol homeostasis (53,54). Furthermore, it has been found that besides ABC-transporters, other, additional membrane-bound transporters, such as uptake transporters from the organic anion transporter protein family and members of the solute-carrier family also participate in drug disposition and safety (55,56). Therefore, it is considered that transporter mediated alterations in drug pharmacokinetics are the result of tightly regulated networks (24), that are constituted of members from distinct families of transporters (57).

Taken together, the debate on the true and entire physiological function of ABC-transporters has not been closed so far. The studies mentioned above illustrate the multifacetted physiological roles of ABC-transporters in the human body and suggest that these transporters are highly complex molecular machines that in essence protect the organsim against various factors derived from the environment.

Role of ABC-Transporters in Disease

ABC-Transporters and Cancer

The last fifty years of cancer research have generated huge amounts of highly useful insights into the genetic and molecular characteristics of this highly complex, multi-facetted disease type (58–64). Furthermore, these research results formed the grounds for the development of many different types of anti-cancer drugs that target a variety of different tumors (65–67). However, shortly after the first successful chemotherapy was applied, it was realized that resistance to these drugs would constitute the major impediment for successful cancer treatment and could negatively influence long-term survival (68). Nowadays it is well known that a large number of patients will develop drug resistance during the course of treatment and will no longer be responsive to other structurally and functionally unrelated anticancer drugs – this phenomenon is known as multidrug resistance (MDR) (69). MDR often results in cancer relapse and is clearly associated with poor overall survival. Extensive research on MDR resulted in the identification of the first human ABC-transporter ABCB1 (P-glycoprotein, Pgp) in the year 1976 by Juliano and Ling (2). It was understood that MDR was the result of reduced intracellular drug accumulation mediated by this membrane-bound, energy-dependent efflux pump (4). ABCB1 is overexpressed in a variety of different haematological malignancies and solid tumours and is characterized by a very broad substrate specificity, which is also referred to as polyspecificity (22,69–76). During the years a large collection of cytotoxic drugs has been identified to be ABCB1 substrates, including Vinca alkaloids, taxanes, etoposide, teniposide, colchicines, actinomycin D, camptothecins, imatinib mesylate, methotrexate and mitoxantrone (77–81). Another interesting finding was that ABCB1-induced MDR could be reversed by verapamil (3). This was the start for three decades of intensive research on inhibitor-based chemosensitization. Verapamil, cyclosporine A, tamoxifen, quinine and many others have been studied for their ABCB1 inhibitory potential (82–84). All of them were characterized by low specificity and side effects due to their primary mode of action. The next generation comprised analogues of the first generation inhibitors, such as dexniguldipine and valspodar, a cyclosporine analogue. These compounds showed critical pharmacokinetic interactions with the concomitant cytostatic comedications due to CYP inhibition and increased hepatic toxicity (85–87). The third generation of inhibitors was designed on basis of extensive SAR studies and a better understanding of how substrates and inhibitors interact with ABCB1 (88–90). This class includes compounds such as tariquidar, zosuquidar, elacridar and laniquidar (21). Many of these have been evaluated in randomized clinical trials. However, so far the principle of ABCB1-inhibition to overcome MDR could not be translated into clinical applications and no drug that reverses MDR has been approved (91–96). A different strategy to overcome MDR, which has gained a lot of popularity in the past, is to deliberately avoid or circumvent ABCB1 interactions. This is

anticipated to be accomplished by rational desigining-out unfavourable chemical scaffolds or substitutions with the aim to deliberately develop ABCB1 non-substrates (97–99). Although many investigations provided some useful strategies (100), many of the novel marketed anticancer drugs, such as the tyrosine kinase inhibitors, still show affinity towards ABCB1 or other transporters (100–102).

In the 90`s of the last century a second ABC-multidrug resistance transporter was identified. The multidrug resistance-associated protein 1 (MRP1 or ABCC1) was found in a small cell lung cancer cell line (6,9). Clinical investigations have shown that ABCC1 levels are elevated in small cell lung cancer (SCLC), non-small cell lung cancer (NSCLC), prostate cancer and that ABCC1 is also associated with accelerated relapse in breast cancer (104–106). Similar to ABCB1, ABCC1 also confers a MDR phenotype and is also notorious for its polyspecific substrate recognition pattern (107,108). Moreover, it shows overlapping substrate specificity with respect to ABCB1 (109–111). In 1998, the third efflux pump ABCG2 (breast cancer resistance protein, BCRP, ABCP, MXR) was discovered (11,112). It is one of the few "half-transporters"- only one transmembrane domain and one nucleotide-binding domain- and therefore needs to form homodimers in order to fulfil its efflux function (113). ABCG2 is also responsible for MDR and strong correlations have been drawn of elevated ABCG2 levels and drug resistance and subsequent disease relapse in various types of tumors, such as non-small cellulcar lung cancer (NSCLC) and (childhood) AML (114–116). The ligand specificity of ABCG2 is partially overlapping with, but also distinct from that of ABCB1 and ABCC1. Additionally, many novel tyrosine kinase inhibitors such as nilotinib, dasatinib and lapatinib are also substrates of ABCG2 (117). In addition to its potential to confer MDR, ABCG2 is also currently evaluated with respect to its protective function in cancer stem cells (see more on this in the next section). In 2010, the first 2D crystals of human ABCG2 were published and subsequent homology modelling elucidated conformational changes that occur upon substrate binding (16).

In addition to the three transporters above, 12 others (ABCA2, ABCA3, ABCB4, ABCB5, ABCB11, ABCC2, ABCC3, ABCC4, ABCC5, ABCC6, ABCC10, ABCC11) have also been linked to MDR in various types of human cancers (118–128).

ABC-Transporters and Cancer Stem Cells - a new paradigm?

Solid tumors, such as breast, colon, lung, prostate and ovary represent a huge therapeutic challenge for contemporary anti-cancer pharmacology (129). Many of these tumors show some kind of therapeutic refractoriness and exhibit a dormant behaviour (130). Intensive investigations on these tumors have shown that besides different subtypes, cells within a given tumor population itself often exhibit distinctive proliferative and differentiative potential (131). This is usually referred to as tumor heterogeneity. This heterogeneity is considered to arise from cancer stem cells (CSCs) (129). These CSCs form the basis for a hierarchical organization of cells within the tumor and they form a subpopulation which is responsible for sustaining tumor growth. CSCs are cells with indefinite self-renewing capacity (in analogy to "classical" hematopoietic stem cells (HSCs)) and have been shown to exhibit

an intrinsic drug resistant phenotype (132). Furthermore, CSCs are considered to contribute to metastasis via mechanisms involving epithelial-mesenchymal transition (EMT) (133). In 2001, Zhou and colleagues described hematopoietic stem cells (HSCs) by their "side population" (SP) phenotype, which is caused by the active efflux of a fluorescent dye by ABCG2 (12). Similar effects were observed by Hirschmann-Jax et al. in 2004, when they analysed cancer cells from 23 neuroblastoma patients (13,134). They found a SP that showed high expression levels of several ABC-transporters (ABCB1, ABCG2, ABCA3) and these CSCs were resistant to mitoxantrone. Another study showed that overexpression of CD133 (a "classical" cancer stem cell marker – notably, ABCB5 is also considered to be a valid biomarker for CSCs in many tumors (135)) in glioma cells induced the expression of ABCB1 and conferred a MDR phenotype to these cells (136,137). Furthermore, Calcagno and colleagues showed that prolonged continuos drug selection of breast cancer cells for drug resistance enriches these cell cultures for CSCs (138). These CSCs appear to be highly resistant to cytostatic pharmacotherapy. This can be directly related to their elevated expression levels of ABC-transporters. A study in lung cancer cells showed that CD133+ CSCs are spared by cisplatin-based chemotherapy (139). Two other studies analysed breast cancer stem cells. They found that residual, posttherapeutic cells are enriched in CSCs (140,141). This was shown for letrozol, paclitaxel and 5-FU therapy. Hence, the link between CSCs and therapeutic resistance has been described as the "axis of evil in the war on cancer" (142). Therefore, it can be concluded that ABC-transporters are even "more than just efflux pumps" (20).

Taken together, these findings provide a potential explanation (on the cellular level) for the origin of MDR and could have intriguing implications for future drug development of anti-cancer agents. Inherent drug-resistant CSC populations are actually – although this seems to be paradoxical at first glance – expanded by drug therapy. Therefore, future efforts should focus on the design of anti-cancer agents that directly target CSCs (this concept is outlined in the following schematic).

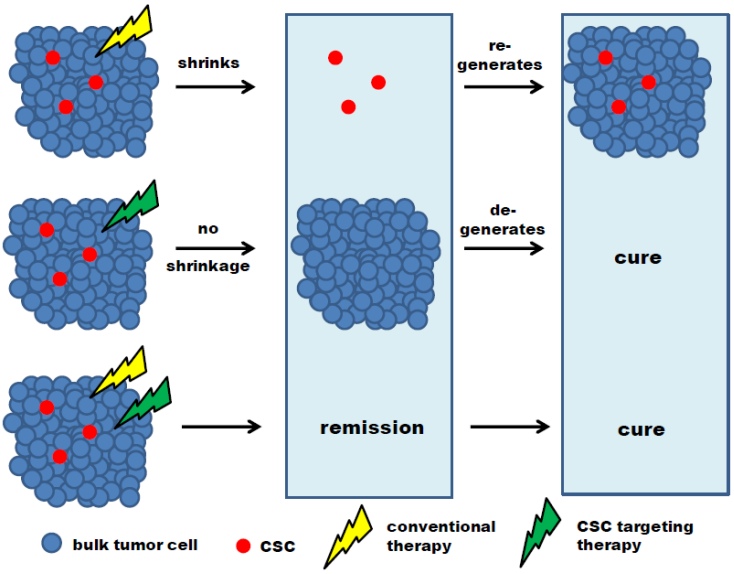


Figure 3: Theoretical concept of a CSC targeting therapy. Conventional therapy is efficient in killing bulk tumor cells, but fails to kill CSCs. Since CSCs are the only cells capable of sustaining tumorigenesis, the tumor regenerates (upper panel). Treatment with a CSC-selective agent initially does not result in any tumor shrinkage. However, the remaining cells are

not capable to sustain tumor volume and eventually the tumor will degenerate (middle panel). The combination of both drug regimens results in cure (lower panel). Adapted and modified from Wang (132).

Recently, Xia and colleagues employed image-based chemical screening to identify 12 potent compounds that effectively kill efflux pump expressing lung cancer cells (143). Interestingly, the identified molecules directly reduced the tumorigenicity of the cancer cells under investigation, which suggests that they affect CSCs. Considering, the role of ABC-transporters in CSCs, suggests that these transporters might represent suitable targets to probably selectively target this cell population which uniquely sustains malignant growth and thereby seems to represent the critical cancer propagating factor (129,144).

ABC-Transporters and other Diseases

ABCB1, the paradigm transporter of the human multidrug transporter family is well known for its dominant contribution to acquired chemotherapy resistance in cancer. However, the extensive studies on this protein have elucidated many other pathological roles of ABCB1. Due to the high ABCB1-expression levels at the blood-brain barrier under physiological conditions, ABCB1 dysfunction is considered to contribute substantially to the onset of neurodegenerative disease, such as Parkinson´s disease (PD) and Alzheimer`s disease (AD) (145–148). Both diseases are age-related and a negative correlation between age and decreased ABCB1 function is established (149). PD is characterized by loss of dopaminergic neurons and the composition of Lewy bodies, which are mainly constituted of α-synuclein. MPP+ a neurotoxin responsible for the onset of PD is also a substrate of ABCB1. According to this, PET radiotracers of ABCB1 substrates could be used to better monitor and to improve diagnosis of PD (150). AD is characterized by progressive cognitive decline, which is a result of protein misfolding and protein aggregation processes mediated by insoluble Aβ-protein. ABCB1 exerts a critical role in Aβ-elimination from the brain and thus may represent a target protein for the prevention and treatment of AD (151). Interestingly, the half-transporter ABCG2 has been shown to have a putative protective role in AD due to inhibition of ROS-mediated inflammation (15). Additionally, an observed localized overexpression of ABCB1 in animal models of epilepsy has given rise to the "multidrug transporter hypothesis" of refractory epilepsy (21,152). The contribution of ABCB1 to epilepsy is mainly based on the following grounds (153–156):

- ABCB1 overexpression seems to be spatially restricted to the seizure region
- This overexpression mainly is associated with pharmacoresistant epilepsy patients
- Most drugs used for the treatment of epilepsy are ABCB1 substrates

Owing to the regionally increased overexpression, the design of specific ABCB1 inhibitors or modulators has been suggested as a useful principle for antiepileptic comedication (156).

Many of the other human ABC-transporters play also significant roles in many other diseases. For instance, ABCC7 (which is also known as CFTR) represents an ion-channel that is responsible for epithelial chloride transport and mutations in this protein form the molecular basis for cystic fibrosis, one of the most common monogeneic diseases (19,157,158). Another member of the ABCC-family, ABCC8 (SUR1), which is part of an

inward rectifying potassium channel, is an important regulator of insulin secretion from pancreatic cells. Notably, ABCC8 is also the primary target of the sulfonylurea antidiabetics. Persistent hyperinsulinemic hypoglycemia, a rare genetic disease which occurs in children results from a defective regulation of insulin secretion and is associated with mutations in the gene encoding ABCC8 (159,160). Other human ABC-transporters also contribute to many monogeneic diseases or complex genetic disorders, such as Tangier disease, Stargardt syndrome, progressive familial intrahepatic cholestasis, pseudoxanthoma elasticum, adrenoleukodystrophy or sitosterolemia (161). Last but not least it needs to be mentioned that not only the human ABC-transporters play important roles in many diseases. Many different bacterial microorganisms also express ABC-transporters as a mechanism of resistance to toxic effects of antibiotics. Consequently, the number of antibiotic resistant bacterial strains is rising and the therapeutic success of antimicrobial drug regimens is decreasing (162–164). Hence, potential "efflux reversers" have shifted into the center of attention in the design of new antibiotics (165).

ABC-Transporters and Drug-Drug-Interactions

Drug interactions are described as pharmacodynamic or pharmakokinetic alterations in response to the concomitant administration of a drug with another substance. This other substance can either be another drug, a component of a certain diet (e.g. grapefruit juice) or a herbal supplement (e.g.. echinacea, St. John`s Wort). It is estimated that 20-30 percent of reported adverse drug reactions are caused by interactions between two or more drugs (69). This incidence rate increases dramatically among older patients and patients subjected to a combination therapy (e.g. cancer therapy or anti-HIV therapy (HAART)). Roughly, drug interactions can be categorized according to their mechanism into either pharmacokinetic, pharmacodynamic or a combination of both mechanisms (71). A well known example for a pharmacokinetic interaction is, when two drugs compete for the same metabolic pathway. As soon as this pathway becomes saturated none of the administered drugs can be fully metabolized, which might result in severe toxicities of the drugs, because of elevated serum concentrations. Pharmacodynamic interactions usually define scenarios when two or more concomitantly given drugs interact with the same molecular target. These additive effects can also result in an unwanted excessive response. Furthermore, opposing pharmacodynamic profiles of drugs can result in unwanted reduced efficacy of these drugs. At this point, it is important to note that drug interactions do not necessarily have to arise in a negative context. In certain cases such effects are intentionally provoked in order to increase the therapeutic benefits of a given treatment (e.g. low-dose treatment with ritonavir in order to boost the antiretroviral effects of its combination partners).

Given the fact that ABC-transporters are found at tissues that either represent barriers for drug bioavailability such as the intestine, or that are responsible for elimination of drugs such as the kidney or the liver, renders these proteins susceptible for drug interactions involving the administration, the distribution and the elimination phases of the pharmacokinetic ADME paradigm.

Interactions with digoxin and other cardiological drugs

A prominent example of such an interaction is the ***digoxin/quinidine*** interaction. In a randomized crossover trial with 6 healthy participants Rameis showed that the concomitant administration of the ABCB1 substrate ***digoxin*** and the ABCB1 inhibitor ***quinidine*** results in a prolongation of digoxin elimination half-lives (166). Fromm and colleagues directly related this interaction to ABCB1. They used mdr1a-knock out mice to confirm that the ABCB1 modulator quinidine reduces the renal secretion of digoxin (167). The digoxin plasma concentrations were not modified in the knock-out models, whereas wild type mice exhibited an elevated digoxin concentration of 73 percent. A similar clinical effect is shown for the protease-inhibitor ***ritonavir***, which is considered to be an (unambiguous) substrate (168). A study with healthy volunteers, who received ritonavir orally and ***digoxin*** via intravenous infusion showed that ritonavir significantly increased the AUC 1.86-fold as a result of renal inhibition of renal digoxin clearance (169).

A study by Drescher et al. has further strengthened the notion of the role of ABCB1 for the elimination of intravenously administered digoxin (170). Two other studies elucidated the interaction of ***digoxin*** together with the antibiotic ***clarithromycin*** (171,172). In a randomized, placebo-controlled double-blind cross-over study involving 12 healthy men it was shown that oral clarithromycin increases the digoxin AUC by 1.7 and reduces the renal clearance of digoxin (172). The effect was more pronounced when both drugs were orally administered. Furthermore, the ***single nucleotid polymorphism*** C3435T of ABCB1 also was shown to influence the serum levels of digoxin. However, regarding this SNP results are contradictory. Some studies show elevated, whereas other studies show reduced concentrations of digoxin (173,174). The influence of SNPs is also of clinical relevance for patients receiving renal allografts. The class of calcineurin inhibitors, which represents the cornerstone therapeutic regimen in preventing allograft rejection, was also associated with lower (but not significant) AUCs in homozygotes with the 1236T allele compared to the wild-type (175).

Other effects were observed in animal studies with the antihypertensive agent ***verapamil*** and the anti-arrhythmic drug ***amiodarone***; both inhibitors of ABCB1 (176). The concomitant administration of one of these two agents with vinblastine (a cytotoxic natural product) resulted in elevated plasma levels of the anti-neoplastic compound. Furthermore, the oral bioavaliability of the beta-blocker ***talinolol*** can be significantly elevated when administered simultaneuosly with the ABCB1-inhibitors ***verapamil*** (177) or ***erythromycin*** (178). The latter interaction was confirmed in a randomized clinical trial involving nine healthy men.

Interactions with proton-pump inhibitors

Additionally, members of the proton pump inhibitors are known to interact with ABCB1 (179,180). A case study reports a severe adverse reaction in a female patient taking ***atorvastatin***, ***esomeprazole*** and ***clarithromycin*** that resulted in rhabdomyolysis. The authors concluded a ABCB1 inhibition of esomeprazole, which resulted in a decreased atorvastatin clearance that led to the rhabdomyolysis (181).

Interplay of ABC-transporters and drug metabolizing enzymes

Hepatic enzymes of the cytochrome-P450 (CYP) family are the main contributors to metabolization processes in the human body. Metabolization can be regarded as a detoxification mechanism that relies upon chemical modification of the substrate. Hence, these proteins are responsible of the ***M*** of the ***ADME***-paradigm (see the figure below). Especially, CYP3A4, a certain isoenzyme, is the main metabolizing agents for approximately 50% of all marketed drugs and therefore critically influences their first-pass kinetics, oral bioavailability and elimination (182). It was found that CYP3A4 is often co-expressed with the ABC-transporter in many tissues (183). Furthermore, there is a remarkable intersection between the substrates and the inhibitors of these two proteins. Additionally, crystal structures have shown that both proteins exhibit large binding cavities to which the chemical and pharmacological diverse ligands can bind in multiple binding modes; sometimes two or more drugs can bind simultaneously to one of the two proteins (14,184). It is also of interest that both proteins are subjected to transcriptional regulation by the constitutive-androstane-receptor (CAR) and the pregnane-X-receptor (PXR), which further supports a functional interaction or relationship (185). Summarizing, both proteins can dramtically influence the pharmacokinetic profiles of various drugs. This plays an important role in the chemotherapy of certain malignancies. In cytotoxic therapy often a multidrug regime is administered to "boost" the effects of the individual agents. However in the gastrointestinal epithelium, CYP3A4 is colocalized with ABCB1 (186). Many commonly used cytotoxic agents, like the ***camptothecins***, ***Vinca-alkaloids*** or ***tubulin inhibitors*** are substrates for both proteins. The interactions between these two proteins have been studied in knock-out mice. Van Waterschoot and colleagues generated single CYP3A4(-), single ABCB1 (-) and dual CYP3A4/ABCB1(-/-) knock-out mice and analysed the systemic drug exposure after oral and intravenous ***docetaxel*** exposure (183). After oral exposure the CYP3A4(-) mice showed an average increase of 11.5-fold of the AUC and the ABCB1(-) mice showed an average increase of 2.8-fold of the AUC compared to the wild-type. Interestingly, the double- mutant CYP3A4/ABCB1(-/-) showed a 72-fold increase of the AUC compared to wild-type. This increase in the bioavailability of docetaxel was also associated with severe toxicities observed in the studied animals. A similar effect was also shown for the dual substrate ***lopinavir***, an antiretroviral agent (183). Another study showed, that the administration of the CYP-inhibitor and ABCB1-inhibitor ***ketoconazole*** led to a serious increase in the bioavailability of the tyrosine-kinase inhibitor ***erlotinib***, whereas the simultaneous treatment with the CYP3A4 inducer rifampicin reduced erlotinib AUC by 66% (71). This synergistic interaction clearly demonstrates the importance of these two detoxifying agents and underlines their contribution on the systemic exposure of oral drugs. Furthermore, it becomes obvious that if one or both of these mechanisms become abrogated, this can lead to detrimental outcomes for patients.

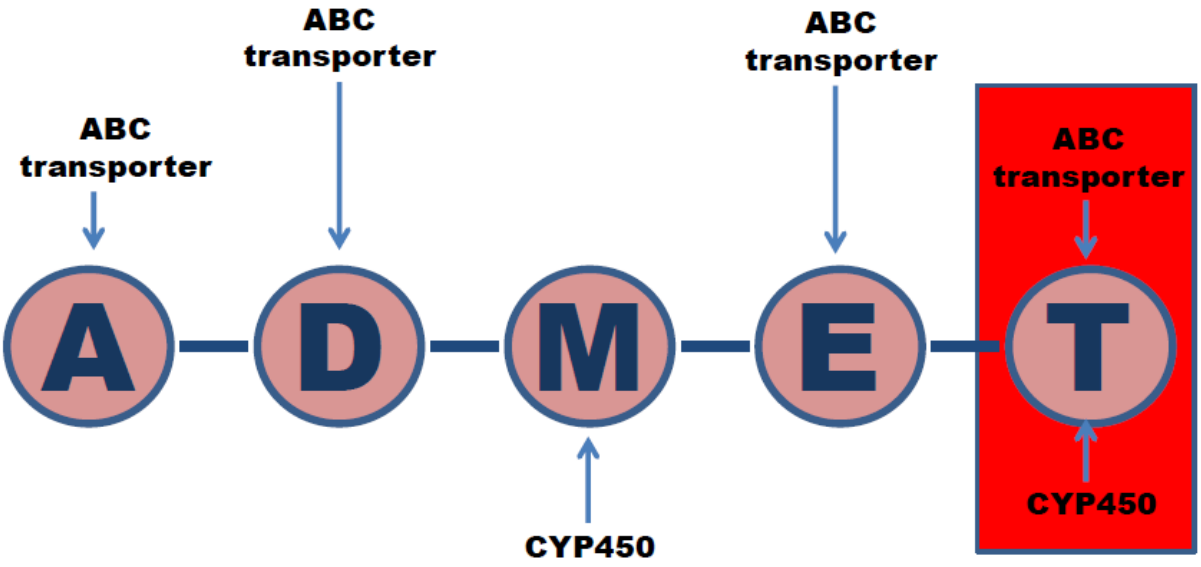


Figure 4: Main contributors to the ADMET paradigm and the interplay and contribution of ABC-transporters and CYP enzymes to the toxicological profile of drugs. Although ABC-transporters and CYP enzymes are different from a biochemical perspective - the former being transporters, which carry their substrates against a steep concentration gradient and the latter being enzymes, which catalyse chemical changes in the structures of their substrates – together they constitute an effective detoxification system, which involves all stages of the ADMET paradigm.

Interactions with the ABCG2 transporter

Additionally, the ABC-transporter ABCG2, which plays an important role in many cancers and is also a contributor to MDR, has been shown to be responsible for drug-drug interactions. A clinical investigation of ***irinotecan***, ***leucovorin***, ***5-FU*** and the EGFR-inhibitor ***gefitinib*** in patients with colorectal carcinoma resulted in increased toxic side effects (187). The observed toxicological profile comprised the haematological and also the gastrointestinal system**.** Both side effect types are characteristic for irinotecan. The side effects could be attributed to an interaction between gefitinib and the camptothecin derivative irinotecan (or its active metabolite SN-38). Irinotecan is a substrate for ABCG2 and gefitinib is an inhibitor of ABCG2. Hence, a putative mechanism for these interaction could be the inhibition of ABCG2 by gefitinib and the subsequent limited the elimination of the substrate irinotecan (188,189).

Chapter 3: Structural and Functional Aspects of ABC-Transporters

The Transport Mechanism of ABC-Transporters

Human multidrug efflux proteins of the ABC-transporter superfamily represent integral membrane proteins that are notorious for their potential to cause multidrug resistance (MDR) and to modulate the pharmacokinetic profile of various systemically administered drugs. ABC transporters utilize the energy provided by binding and hydrolysis of ATP to actively extrude xenobiotic compounds out of living cells. The human genome encodes for 48 ABC efflux pumps, which are grouped into seven subfamilies (ABC-A to ABC-G) according to their primary sequence. Although there are major structural differences among these seven families, all 48 transporters are constituted of four functional and structural similar core domains. These four core domains, which form the minimal functional unit, are:

- **Two transmembrane domains (TMDs):** each TMD consists of multiple membrane-spanning α-helices, which constitute together the binding sites and form the translocation pathway for substrates.
- **Two nucleotide binding domains (NBDs):** NBDs are required to confer conformational changes that are evoked by nucleotide binding, hydrolysis and ADP + Pi release

The NBDs of all ABC transporters (whether prokaryotic or eukaryotic) are highly homologues to each other, whereas major differences in the sequences are observed in the TMDs.

Extensive biochemical and structural inverstigations (190) including photoaffinity labeling (90), site-specific mutagenesis (90,97,191–193), Cys-scanning (194,195), utilization of pharmacological chaperones (196), nucleotide trapping (197,198) as well as electron microscopic and crystallographic studies (14,197,199–202) have elucidated the highly complex transport cycle that is used by ABC transporters to facilitate extrusion of xenotoxins out of cells against a concentration gradient. This transport cycle embodies mainly four steps and is schematically depicted in the following figure:

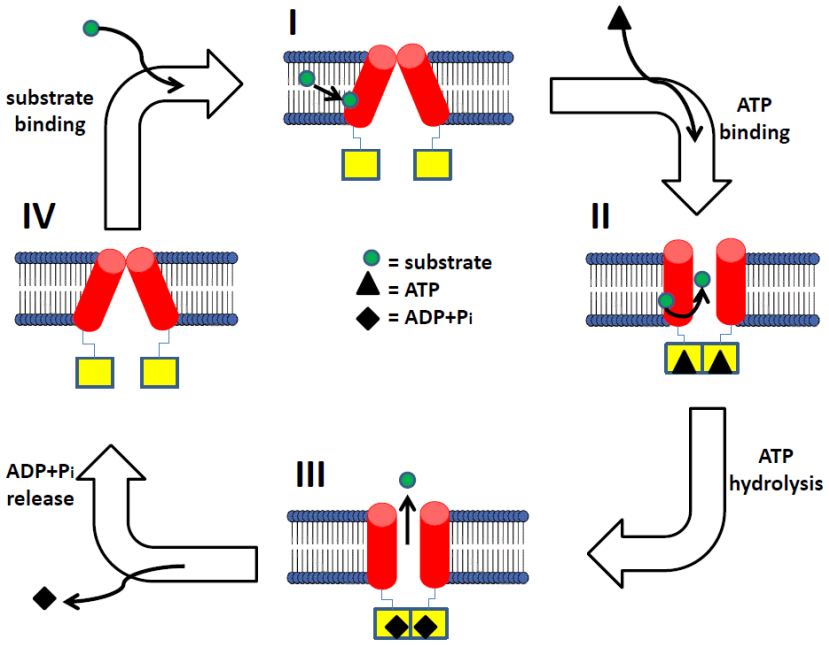


Figure 5: The transport cycle of ABC-transporters.

- **Step I: Substrate binding induces conformational changes in the NBDs.** The first step in the catalytic transport cycle includes the binding of substrate to its high-affinity binding site at the TMDs. For some substrates it has been shown experimentally that binding occurs at the TMD/TMD interface (90,203–206). This binding induces conformational changes in the NDBs, which result in lowering the required activation energy for ATP-dependent dimerization of the NBDs (207–211).
- **Step II: ATP binding induces conformational changes in the TMDs.** In the second step the binding of two molecules of ATP provides the necessary power stroke for switching the high-affinity substrate binding site in the TMDs to a low-affinity substrate binding site, which is now accessible from extracellular regions (88,89,212,213).
- **Step III: ATP hydrolysis.** The hydrolysis of ATP to ADP + Pi and the subsequent release of bound substrate to the exterior results in a destabilization of the closed NBD dimer and initiates resetting of the transporter to its basal conformation (214,215). The initial trigger is not known; however two main hypotheses claim: (I) release of substrate mediates conformational changes that are transmitted to the NBDs; (II) autocatalytic ATP hydrolysis (216,217).
- **Step IV: release of Pi + ADP to restore the basal state.** After hydrolysis of ATP electrostatic repulsion destabilizes the closed dimer and a subsequent rigid body movement leads to the sequential release of Pi and ADP (218–220).

The Polyspecific Ligand Recognition of ABC-Transporters

The primary purpose of ABC-transporters is suggested to be the protection of the organism from exogenous toxins by active extrusion of these out of living cells. In order to properly execute this function it is necessary that these proteins are able to recognize and bind a wide variety of chemically and functionally unrelated molecules. However, the molecular basis for

this polyspecific ligand recognition pattern still remains unsolved. A large body of early published biochemical and pharmacological data revealed one recurrent theme that is nowadays considered to be the least common denominator of ligand binding to ABC-transporters: *hydrophobic ligands do interact with binding sites in ABCB1 that lie within the membrane* (203,204,221,222). This has been exemplified by Friche et al., who showed that the potency of inhibitors to interfere with the transport of anthracyclines out of cells by ABCB1 is directly proportional to their ability to partition in the lipid bilayer of the membrane (223). Furthermore, the presence of at least four different, partially overlapping but still distinct solute binding sites was detected using kinetic assays (88,89). At least three different sites for the substrates vinblastine, paclitaxel and rhodamine123 and one modulator site have been detected. The fourth site is also the binding site of the inhibitors elacridar and nicardipine (21). These findings strongly support the notion that ABCB1 accomodates different drug binding sites into one huge binding cavity which could be as large as approximately 6000 $Å^3$ (14). Additionally, the publication of various X-ray structures of bacterial homologues as well as the structure of murine ABCB1 revealed highly interesting insights into the functional and biochemical features of these transporters (14,202,224,225). These structures were subsequently subjected to homology modelling to project the 3D-conformation of the bacterial structures onto the sequence of human transporters. These efforts and results have been reviewed by several authors (98,226). The available X-ray structures of several ABC-transporters are summarized in the table below.

Table 1: Summary of available X-ray structures of bacterial or mammalian ABC-transporters.

Species	Protein	Conformation	Nucleotide	Ligand	PDB	Resolution (Å)
bact.	Sav1866	Outward	ADP	none	2HYD	3
bact.	Sav1866	Outward	AMP-PNP	none	2ONJ	3.4
bact.	MsbA	Inward	none	none	3B5W	5.3
bact.	MsbA	Inward	none	none	3B5X	5.5
bact.	MsbA	Outward	AMP-PNP	none	3B5Y	4.5
bact.	MsbA	Outward	ADP-OV	none	3B5Z	4.2
bact.	MsbA	Outward	AMP-PNP	none	3B60	3.7
mamm.	ABCB1 (murine)	Inward	none	none	3G5U	3.8
mamm.	ABCB1 (murine)	Inward	none	QZ59-RRR	3G60	4.4
mamm.	ABCB1 (murine)	Inward	none	2xQZ59-SSS	3G61	4.35

Especially, the structures of murine ABCB1 have attracted much attention since it shows 87% sequence identity to the human protein. Additionally, the structures, which resemble the inward conformation are also co-crystallized with two hexapeptide ligands. Probably the most interesting finding of this publication was that ABCB1 is able to accommodate two ligands simultaneously. This highlights that ABCB1 can adopt different side chain conformations of transmembrane amino acid residues to create binding sites for different ligands. Another recent biochemical investigation using photoaffinity labelling of propafenone analogs showed that ABCB1 is able to provide at least two different translocation pathways for its solutes and that certain substrates make use of one pathway while others prefer the different one (227).

These recent insights aid in the understanding of the quantitative structure activity relationships of drugs interacting with ABCB1.

Categorization/Definition of Compounds interacting with ABC-Transporters

The categorization of the type of interaction of ABC-transporters with drugs is a key step in the determination of the pharmacological activity and the pharmacokinetic profile of these drugs. On principle certain types of interactions of molecules can be distinguished, which can be broadly divided into direct and indirect interactions:

Direct interactions:

- **Substrate –** molecule is effluxed out of living cells by the transporter
- **Inhibitor –** molecule directly interferes with the transport capacity; inhibits transport of a substrate; in case the substrate is a cytotoxic agent, MDR-inhibitors might restore the inherent cytotoxic potential of the substrate through the inhibition of efflux. Therefore, inhibitors are often characterized as modulators (see below).
- **Modulator –** behaves on a "macroscopic" level like an inhibitor (i.e. is able to reverse a MDR phenotype and abrogates the transport of a reference substrate), but is by itself transported; hence the observed effect is the result of direct competition for a drug-binding site by two substrates ("*competitive substrate*"). This type of interaction is especially delicate, since such compounds might be reported as substrates in one type of assay (ATPase assay), but will be categorized as inhibitors in another type of (Calcein-AM) assay (see below and section C).

Indirect interactions:

- **Inducer –** molecule that upregulates transporter expression; some substrates have been shown to upregulate the expression of the respective efflux pump via binding to nuclear receptors (228).
- **MDR-reverser –** restores the cytotoxic potential of a reference drug; often a direct inhibitor of the transporter, but can also interfere with transporter expression levels. Notably, the term "reverser" is often used interchangeably with "modulator" in the literature.
- **MDR-selective molecule –** is more cytotoxic in MDR/transporter-overexpressing cells than in cells that do not show a MDR phenotype/do not express the respective transporter. This hypersensitivity of resistant cancer cells to certain compounds is also termed "collateral sensitivity". Although this paradoxical mechanism of cytotoxicity is not very well elucidated on the molecular level (and it must be speculated that the effects of these agents do not result from an interaction with ABC-transporters), it still represents a promising new strategy to improve response to chemotherapy. "MDR-selective" molecules were acknowledged only recently (23,229–232), however this phenomenon was known for quite a long time and has been described for drugs, such as verapamil, nicardipine, diltiazem and colchicine (233–237).

Assays measuring direct interactions:

On principle, there are several assays available that can be applied to categorize the type of interaction a compound exhibits with an ABC-transporter. These three assays are comprised of the monolayer efflux assay, the ATPase activity assay and the calcein-AM fluorescence assay. A detailed overview of these assays is given in Part C of this thesis. Briefly, the monolayer assay is considered to represent the current state-of-the-art method to determine whether a compound is an ABC-transporter substrate. It measures the ratio of basolateral-to-apical (B→A) permeability versus the apical-to-basolateral permeability (A→B) of a compound. Unfortunately, this type of assay is very labour intensive and is unsuitable for high throughput applications. The ATPase assay which assesses the ratio between release of inorganic phosphate from ATP in the presence of the molecule of interest and its absence, is more suitable for high throughput applications, but is not able to differentiate between substrates and inhibitors. The third type of assay can only be used to assess if a compound is an inhibitor. However, its application might be feasible to exlude ambiguous results from the other two assays. A summary of these assays and a potential categorization scheme is given in the following table. It is necessary to mention that only a combination of these assays can unravel the full interation pattern of a molecule. Lessons learnt from previous studies highlight that many molecules might be assigned differently in dependence of the assay used. A prominent example of this *"categorization controversy"* are the different findings observed for the Ca-channel blocker verapamil (168):

- Tsuruo showed in 1981 that verapamil can reverse the MDR phenotype (3).
  - this categorizes verapamil as a ***MDR-reverser*** (eventually an ABCB1-inhibitor)
- Warr showed in 1986 that MDR-CHO cells were hypersensitive to verapamil (235)
  - this categorizes verapamil as ***MDR-selective agent***
- Tiberghien showed in 1996 that verapamil is not transported in an efflux assay (238)
  - this categorizes verapamil as a ABCB1 ***non-substrate***
- Litman et al. showed in 1997 that verapamil strongly increases ATPase activity (239)
  - this led the authors to the conclusion that verapamil is a ABCB1 ***substrate***
- Pauli-Magnus et al. showed that verapamil enhances calcein fluorescence (240)
  - this categorizes verapamil as an ABCB1 ***inhibitor***

Another example is the anti-retroviral ritonavir; depending on the literature source it is categorized either as substrate or inhibitor (168,169). In order to account for this categorization problem Polli et al. proposed that only a combination of multiple assays can be used to discriminate between: *unambiguous substrates* (e.g. loratadine, loperamide, quinidine, ritonavir), *transported substrates*, *nontransprted substrates* (e.g. ketoconazole, midazolam, verapamil), *unambiguous nonsubstrates* (e.g. amantadine, lidocaine) *and inhibitors* (e.g. GF120918). However, it needs to be stated that such a detailed categorization and combination of assays is only available for few compounds and it can be concluded that most of the literature data might not unravel the full picture of interaction for some compounds.

Table 2: Summary of classical assays used to categorize the interaction type with ABC-transporters. Adapted from (168)

Property	**Assay**		
	Efflux assay	**ATPase assay**	**Calcein-AM assay**
Used for	Substrates	Substrates	Inhibitors
Can distinguish between (Substrates/Inhibitors)	Yes	No	No
Analytical technique	Spectroscopy	Absorbance	Fluorescence
Throughput	12 cpds./week	150 cpds./week	200 cpds./week
Threshold for activity	BA/AB ratio > 2.0	ATPase ratio > 2.0	>10% of reference[a]
Classification scheme proposed by Polli et al. (168)**:**			
Unambiguous substrate	Yes	Yes	Yes
Transported substrate	Yes	Any	Any
Nontransported substrate	No	Yes	Yes
Unambiguous nonsubstrate	No	No	No
Inhibitor	No	No	Yes

Additionally, to the categorization scheme presented above, the International Transporter Consortium (ITC) presented a decision tree for ABC-transporter substrates in 2010 (24):

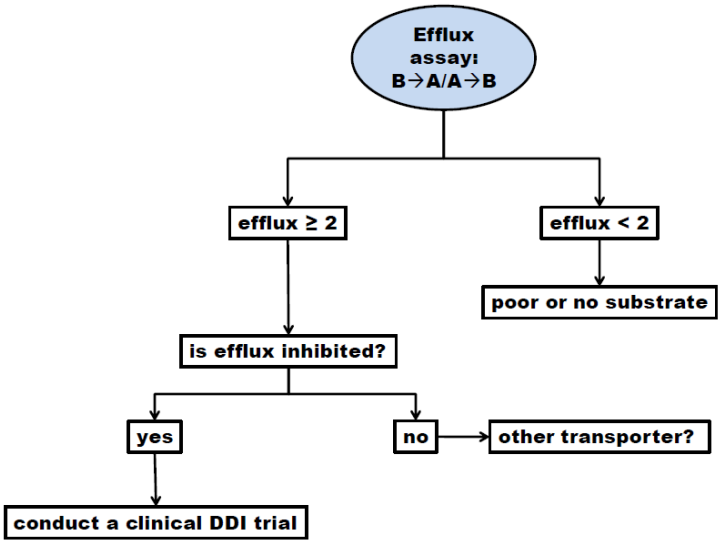


Figure 6: Decision tree proposed by the ITC for ABC-transporter substrates. DDI= drug-drug interaction. Adapted and modified from (24).

Furthermore, the authors proposed that in case a potential ABC-transporter substrate shows a ratio of the concentration of the molecule in the gut lumen and its IC_{50} for inhibition of higher than 10, a clinical drug-drug interaction trial should be conducted. They propose digoxin as a suitable *in-vivo* probe substrate for such a clinical investigation. However, the relevance of this approach was critizised by other authors (241). Similar categorization schemes have also been provided by the ITC for inhibitors and other transporter families in the same publication.

Surrogate (cytotoxicity) assays for substrates *and/or MDR-selective agents:*

Additionally, to the three assays mentioned above, substrates of transporters can also be measured using surrogate *in-vitro* endpoints. In other words, the interaction between the compound and the transporter is not directly assessed, but is indirectly inferred from the conceptual design of the assay. Such surrogate endpoints have been widely used for cytotoxic compounds. The underlying principle of such an assay is to compare the cytotoxicity of a compound in a transporter overexpressing cell line to its cytotoxicity in the corresponding maternal cell line that does not express the respective transporter. This can be easily done by simply calculating a resistance ratio (100). Since the only difference between the two determined activities is the presence/absence of the transporter, it can be concluded that the observed effect is a direct consequence of the active extrusion of the substrate by the efflux pump.

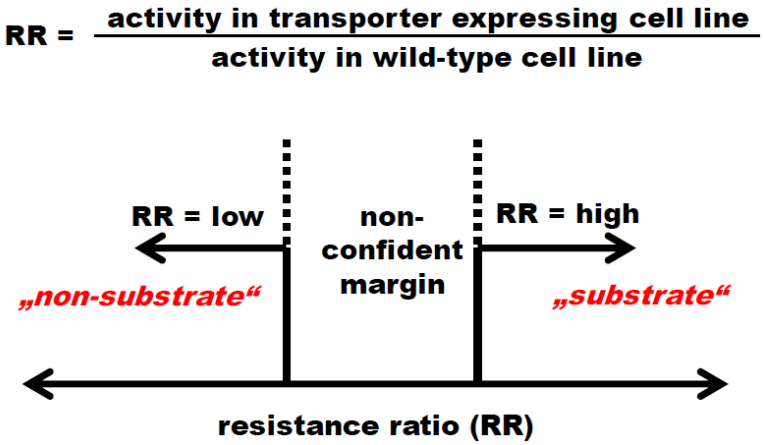


Figure 7: Theoretical concept of the resistance ratio, a surrogate endpoint for the categorization of ABC-transporter substrates and non-substrates.

The potential advantage of this endpoint is that it can be easily implemented in medium to high throughput screening projects, but the disadvantage is that it can only be applied for compounds displaying intrinsic cytotoxicity. One data set, provided by Boehringer Ingelheim Austria that is used in this thesis makes use of this surrogate endpoint in a slightly modified manner.

Another, but slightly more complex surrogate measure has been applied by Szakacs et al. (242). In this study the authors used quantitative real-time PCR using SYBR Green chemistry to elucidate the expression levels of all 48 human ABC transporters available in the NCI60 panel. Afterwards, they calculated Pearson`s correlation coefficients between the mRNA levels in the 60 cell lines and the growth inhibitory concentration of more than 1400 drug candidates. This database allows the categorization of three classes of compounds (see schematic below). Those molecules that show a negative correlation can be considered to be substrates (i.e. cytotoxic potential decreases with increasing transporter mRNA levels). Those molecules that do not show any correlation can be considered to be non-substrates (i.e. their cytotoxicity is not modified by the concentration of transporter mRNA levels). Those molecules that show a positive correlation are considered to be MDR-selective compounds (i.e. their cytotoxicity is potentiated in the presence of increasing transporter mRNA levels).

The potential drawbacks of this database are that it is based on two assumptions:

- mRNA levels are valid surrogates for functional ABC-transporters
- the Pearson correlation as a bivariate measure of tendency is extremely sensitive to outliers

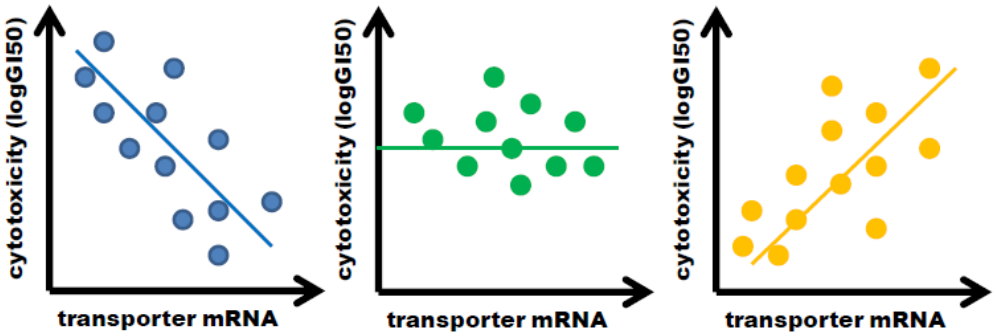


Figure 8: Simplified interpretation of the NCI60 drug sensitivity profiling by Szakacs et al. Three classes of compounds can be identified from this screen on basis of their correlation between cytotoxicity and transporter expression in 60 different cancer cell lines; note: each dote represents one cell line

Nevertheless, it needs to be mentioned that this data set represents the largest public available set with information on the relationship between more than 1400 compounds and all 48 human ABC-transporters. Furthermore, this database does not suffer from intralaboratory differences like other data sets that are compiled from different literature sources. In this thesis this data set is used several times for modelling purposes.

In this thesis data sets comprising all the different types of assays described above are used to characterize the ligands of ABC-transporters.

PART B: CONCEPTUAL & METHODOLOGICAL FRAMEWORK

Chapter 4: An Introduction to Pharmacoinformatics

The development of new drugs is a long and exhaustive process that embodies mainly two different, but often interconnected phases. The first phase is described as the "Research"-phase, whereas the second phase is referred to as the "Development"-phase. Together, these two phases constitute the "R&D"-paradigm of drug discovery and development (243). The "R"-phase embodies many steps that include in the sequence of events: *target identification (TID) – target validation (TV) – lead identification (LID) – lead optimization (LO)*. These steps integrate the whole experimentally determined spectrum of medicinal chemistry. The "D" phase embodies the *preclinical* (in-vivo, but not in-human) and the *clinical* (evaluation in human subjects) stages (244). Considering the fact that on the one hand financial costs are steadily rising during the "R&D" of a novel investigational product, whereas on the other hand the amount of pharmacological data is rapidly expanding, the use of information technologies and the application of sophisticated informatics has gained more and more popularity in many pharmaceutical companies and also in academic research groups in the last decades. The use of these technologies serves many different purposes, such as data organization and data warehousing, gathering and synthesizing already existing research results and eventually predicting and profiling the pharmacological fate of new drug candidates. These methodologies are primarily applied with the intention to reduce costs and eventually to prevent attrition or withdrawal of marketed drugs or candidates which have already advanced through many stages of clinical devlopment.

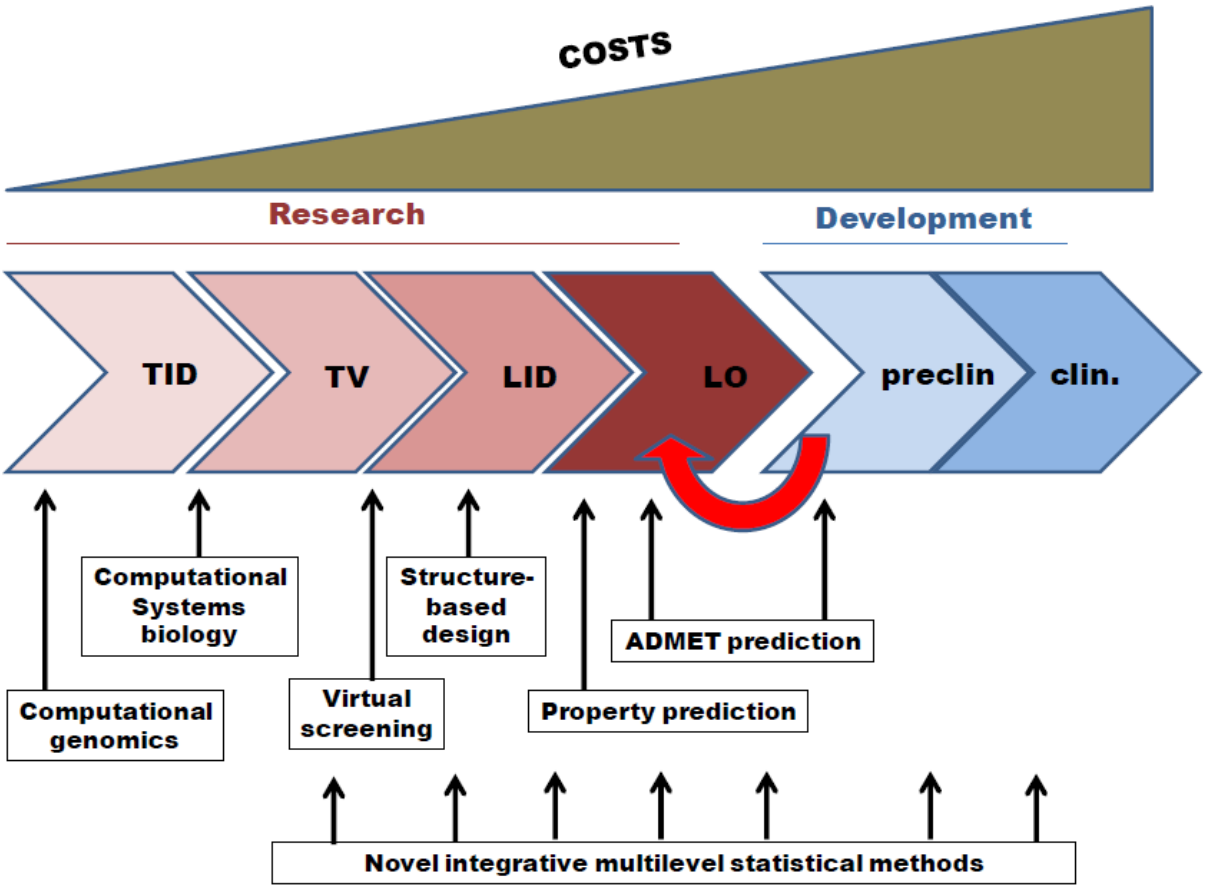


Figure 9: Pharmacoinformatics methods are an integral part of the Pharma R&D process.

Hence, pharmacoinformatics can be viewed as a scientifc discipline, which aims to provide a deeper understanding of the medicinal chemistry of medium- or large-scale data sets using computational (*in-silico*) methods. Therefore, pharmacoinformatics represents the synergy of classical, experimental medicinal chemistry and information technology and is nowadays integrated in many stages of pharmaceutical research (see schematic above). Depending on the particular research project pharmacoinformatic methods can be divided into two broad classes:

- **Target/Structure-based methods**
- **Ligand-based methods**

Target- or structure-based methods often represent means to analyse and interpret putative drug targets (e.g.: computational genomis or systems biology) or directly make use of the presence of a known target (e.g.: protein) structure (e.g.: docking in the context of virtual screening (245,246)). The target structure is usually determined experimentally using X-ray diffraction or protein NMR methods (247,248). However, recent advances in the field also enable structure-based drug design without the explicit knowledge of the target protein (e.g.: homology modelling, de-novo protein modelling (226,249–252) or pseudo-receptor modelling (253)). Homology modelling is widely used in the field of ABC-transporters, since a reliable, high-resolution crystal structure is not available for any human transporter. A comprehensive overview on these homology models is given in Klepsch et al. and Ravna et al. (98,99,226).

Contrary to that ligand-based methods do not require the existence of a known target structure, but are dependent on data sets of ligands that share a given property under investigation (e.g.: affinity to the target, blood-brain barrier permeability, substrates of ABC-transporters,…). Hence, ligand-based methods depend on the accurate and precise measurement of the biological/pharmacological variable of interest. All these methods rely upon the same principle, that molecular structure of ligands can be directly related to molecular/pharmacological effects. Ligand-based methods include:

- **“classical” QSAR analysis**
- **Similarity Searching and pharmacophore modelling**
- **Machine learning modelling of target- or property-based characteristics of chemicals**

“Classical” QSAR analysis dates back to the days of Crum-Brown and Fraser, Meyer and Overton, Hammett and Corvin Hansch (254,255). It is usually a very powerful method, that utilizes more or less sophisticated versions of multiple linear regression. However, the main drawback of “classical” QSAR is, that it expects an identical mode of action within the compounds in the data set and that it is usually only applicable for small-sized, homogenous chemical series. Contrary to classical QSAR modelling, pharmacophores describe the minimal spatial localisation and distribution of functional groups or physicochemical features of a ligand set (256). Such models are often used in the context of scaffold hopping and also can serve as inputs for similarity searching experiments (257). Pharmacophores are always dependent on accurate three dimensional ligand structures and therefore their calculation can

be computationally expensive and time consuming. Machine Learning (ML) methods embody a very versatile and highly multi-facetted means for ligand-based drug design and they are extensively used and evaluated in certain stages of the drug discovery process (258–260). In this thesis the majority of the findings resulted from the application of ML algorithms. Therefore, this family of methods shall be explained in more detail below.

Chapter 5: Machine Learning as Tool for Pharmacoinformatic Studies in Ligand-based Drug Design

An Overview of Machine Learning

This chapter aims to provide the reader with the basic fundamentals of Machine Learning (ML) theory. Furthermore, it describes some commonly used algorithms. The main objective is to highlight the underlying principles of the applied algorithms in an *easy-to-read* manner. A more *in-depth* description of the applied algorithms is given in Part D of this thesis.

Components of Machine Learning Models

In its simplest form a ML model is constituted in a four-step procedure. The first step includes data collection and descriptor calculation. In the context of pharmacoinformatics data collection embodies the retrieval of chemical structures of interest and also includes annotating the biological response (with respect to one or multiple targets) accompanying these structures. The response variable (i.e. the biological effect under investigation) can be either categorical (e.g.: +1 or -1; active or inactive) or continuous (e.g.: K_i or IC_{50} values). Additionally, for each instance (i.e.: molecule) a variety of chemical descriptors can be calculated that serve as numeric representations of the chemical information inherent to the molecules under investigation (261). At this stage the input for a ML model, the data matrix, is generated. Nowadays, a variety of different chemical descriptors are available, that encode molecules at different levels of complexity and sophistication. Roughly, chemical descriptors can be categorized into three groups:

- **1-dimensional**: usually computed from the elemental formula alone (e.g.: atom-type counts),
- **2-dimensional**: additionally consider the molecular constitution (e.g.: connectivity of atoms),
- **3-dimensional**: represent the most complex types of descriptors; that account for the 3D strucutre of molecules and thereby consider spatial localization of molecular features and/or shape.

In the second step various types of data preprocessing can be conducted (262). Usually this step is optional and thus is not a requirement of a ML model. However, in the case of highly complex modelling tasks, this step is highly recommended. Data preprocessing usually includes various forms of dimensionality reduction methods, which represent linear combinations of the original data with the aim to retrieve low-dimensional representations. Alternatively, various feature selection algorithms can be applied at this stage. Feature selection (FS) describes the process of selecting the most useful/informative descriptors out of a large pool of initially calculated ones. It is important to note, that dimensionality reduction produces a new input set due to mathematical transformations, whereas FS retains a smaller subset of the original descriptors (263).

Next, in the context of data mining a ML algorithm is inferred, that relates the input data matrix to the ***Y***-variable in a mathematical manner. In analogy to the variety of chemical descriptors that are available nowadays, also a vast compilation of ML algorithms is available.

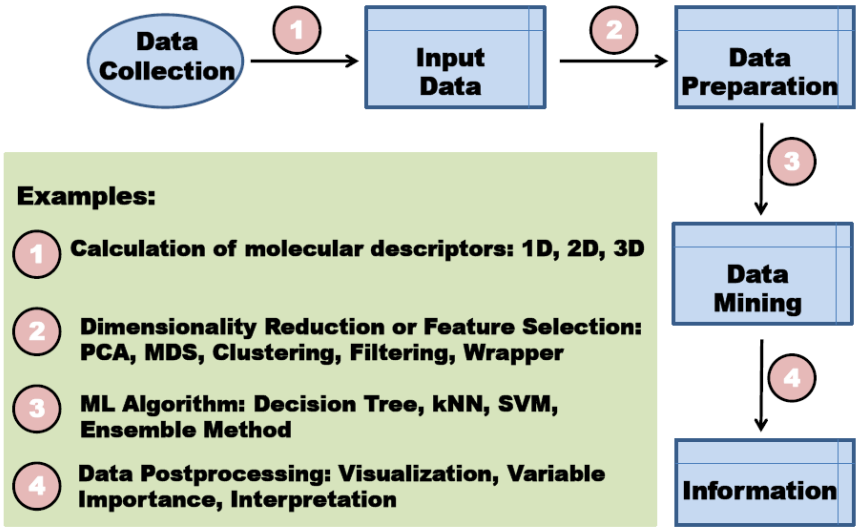


Figure 10: Components of machine leaning

These algorithms differ remarkably with respect to their predictive performance, their calculation speed and also with respect to their capabilities to provide interpretation for the user. Furthermore, it is also of importance to consider the individual algorithmic capabilites to handle large data sets, to deal with redundant attributes and the ease of parameter handling in the selection of the algorithm. Definitely, there is no *universal* algorithm that performs best for each modelling problem. However, benchmarking studies have shown that so-called *ensemble methods* (Bagging, Random Forests, Boosting) outperform classical single-model learning schemes (264). Among the class of *single-learners*, Support Vector Machines and Naive Bayes algorithms usually give better performances than simpler algorithms, like *k*NN and decision trees.

Finally, the ML model is analyzed in the course of data postprocessing, which primarily aims to retrieve useful information out of the learned model. Here, a variety of visualization methods (like radial visualization plots, partial dependence plots, or MDS plots) can be applied. Additionally, methods to assess variable importances can be used to interpret the model.

Conventions in Machine Larning

Like many other scientific disciplines, the field of ML uses some common conventions. As outlined in the section above the input of a ML model is a data matrix. By convention this data matrix $\boldsymbol{X}=\{\boldsymbol{X}_{ij}\}$ is defined by $\boldsymbol{m}$ rows (instances) and $\boldsymbol{n}$ columns (properties/predictors or features). In the context of molecular modelling the $\boldsymbol{m}$ rows ($\boldsymbol{X}_i$) represent the $\boldsymbol{m}$ molecules. The $\boldsymbol{n}$ columns ($\boldsymbol{X}_j$) on the other hand represent properties of each instance or in other words the columns contain the calculated descriptor values for each compound. Additionally, in the case of supervised learning a $\boldsymbol{Y}$-vector is also submitted to the ML algorithm. The Y-vector is of length $\boldsymbol{j}$. In the case of classification this $\boldsymbol{Y}$-vector contains the biological/or pharmacological empirical measurements in a binnarized form.

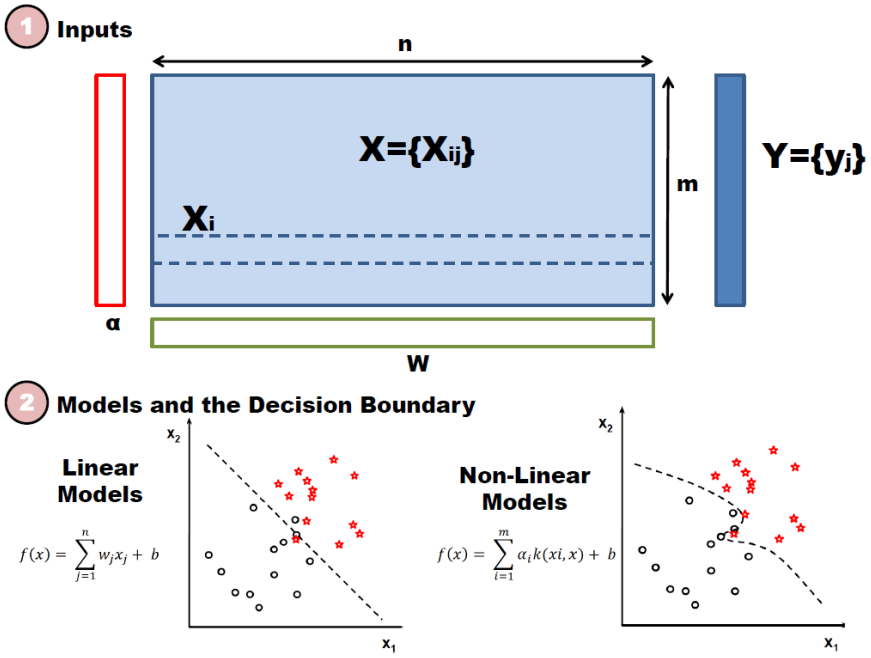


Figure 11: Conventions in machine learning

Since it is the goal of ML to relate the input matrix $\boldsymbol{X}=\{X_{ij}\}$ to the response vector $\boldsymbol{y}=\{y_j\}$ in a mathematical manner, weights are assigned to the matrix $\boldsymbol{X}$ to accomplish this task. In dependence of the algorithm used these weigths can be assigned along the columns $\boldsymbol{x_j}$ and are by convention termed $\boldsymbol{w}$ or can be assigned along the instances $\boldsymbol{x_i}$ and are termed $\boldsymbol{\alpha}$. In case of classification, the mathematical relation between $\boldsymbol{X}$ and $\boldsymbol{y}$ that is inferred in the course of ML aims to derive some type of decision boundary (a mathematical function on which the claasification decision is based). The underlying principle of the used algorithm dictates if the decision boundary is of a linear type (e.g.: multiple linear regression, PLS) or if it is non-linear (e.g.: decision trees, Kernel Methods) and additionally if $\boldsymbol{w}$`s or if $\boldsymbol{\alpha}$`s are used for computation. It is important to note that the term *linearity* in *linear methods* refers to both *linearity* with respect to $\boldsymbol{X}$ and also to *linearity* with respect to the parameters the model is based on. Non-linear models do not require *linearity* in the input matrix $\boldsymbol{X}$, but might require *linearity* with respect to the parameters they are based on (265). In the case, that $\boldsymbol{X}$ does not show *linearity*, feature transformations such as Principle Component Analysis (PCA) are necessary in the course of data preprocessing to correctly apply linear ML algorithms. Usually, non-linear models outperform linear methods with respect to performance, but are often more difficult to interpret (266). Nevertheles, non-linear methods often require the application of sophisticated post-processing protocols because they are usually much more complex and therefor difficult to interpret.

Types of Machine Learning

As outlined in the section above ML models aim to relate $\boldsymbol{X}$ to $\boldsymbol{y}$. This type of relation can be used as one of several ways to categorize the different ML algorithms into two large groups (265). To explain the term *supervised*, the following analogy can be drawn:

- **supervised learning = learning with a teacher**. In this metaphor the ML model (student) presents an answer $\boldsymbol{y_i}$ for each $\boldsymbol{x_i}$ in the data set and the supervisor (teacher)

provides the correct answer. In other words, in the case of supervised learning, the ***X*** AND ***y*** are used by the algorithm to compute a decision boundary. Algorithms that are supervised by definition are: *k*NN, SVM and Naive Bayes.

- **unsupervised learning = learning without a teacher**. In this scenario one has a data set ***X*** and it is the goal to directly infer the class membership without making use of ***y*** (without the help of a supervisor or teacher). Commonly used unsupervised algorithms are: principal component analysis (PCA), multidimensional scaling (MDS), Clustering or Self-Organizing Maps (SOMs). It is important to note that some supervised algorithms can also be used in an unsupervised setting (e.g.: Random Forests).

A clear disadvantage of unsupervised methods is that the quality of the retrieved models is usually very difficult to estimate. In case of supervised learning clear measures of success such as the Matthew`s correlation coefficient in case of classification or the r^2-value in regression do exist. In order to account for this, unsupervised methods often use visualization techniques to estimate performance (e.g.: PCA and SOM). An unsupervised algorithm that is often used throughout this thesis for visualization of data sets is the Principal Component Analysis, which is explained in more detail below.

Principal Component Analysis (PCA). PCA is an unsupervised pattern recognition technique that is commonly applied during the course of data pre-processing. The primary objective of PCA is to derive a set of *new, transformed* features (called *principal components (PCs)*) of the original matrix ***X***. These PCs represent a *low-rank approximation* of ***X*** and hence can be used for the visualization of high-dimensional data in 2D or 3D. According to Jolliffe PCs meet the following criteria (267,268):

- they are linear combinations of each other,
- they are perpendicular to each other,
- and they capture the maximum amount of information in the data.

The underlying principle of a PCA comes from the field of matrix algebra, more specifically relies on the calculation and properties of eigenvectors and their cognate eigenvalues. An important prerequisite for the calculation of eigenvectors and eigenvalues is that they can only be computed for *squared matrices*. Thus, the usually *n x m*-matrix ***X*** needs to be transformed into a n x n-matrix prior to calculation. In many cases the covariance-matrix *Cov* is used to derive this squared matrix, although other calculations such as the correlation matrix *Cor* are also possible. Eigenvectors and eigenvalues are then defined in the following way:

$$\boldsymbol{A\nu = \lambda\nu}$$

Equation 1: Eigenvector decomposition

The above equation is interpreted as folllows: The vector ν is an *eigenvector* of the squared matrix ***A*** (which is for example the covariance matrix of ***X***), if there is a scalar termed λ corresponding to ν. An important feature of eigenvectors is that there are *n* of them for a *n x n* matrix. Furthermore, eigenvectors and eigenvalues always appear in pairs and the magnitude of the eigenvalue corresponds to the amount of information captured by the corresponding eigenvector. Thus, eigenvectors can be ordered with respect to decreasing values of their eigenvalues. A plot of eigenvalues calculated by PCA is called the *scree-plot*. The eigenvectors constitute the new, orthogonal coordinate system. It is important to note that

eigenvectors are computed in such a way, that they maximize the amount of information (variance) in the new coordinate system.

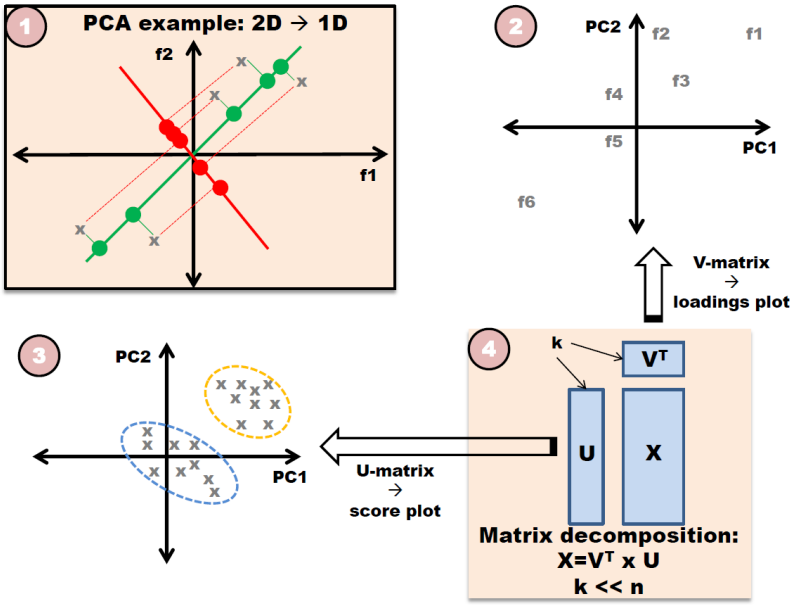


Figure 12: Dimensionality Reduction via Principal Component Analysis. 1) depicts a schematic illustration of dimensionality reduction via PCA; in a 2D space a new 1D coordinate system is established by capturing the maximum amount of variation within the data. The green axis fulfils this criterion, whereas the hypothetical red axis captures a much smaller amount of information since the data points (red) are clustered tightly together. Hence, the green axis is the preferred one in this scenario. 2) - 4) graphically visualize the different types of information that can be retrieved from a PCA matrix decomposition (shown in 4). From the U-matrix the score plot is derived and from the V-matrix the loading plot is derived. NOTE: the score plot visualizes the instances, whereas the loading plot visualizes the features from the original data set.

The first two or three eigenvectors can be plotted in a scatter plot and hence mirror the above mentioned *low-rank visualization* of the original data set. The scatter plot of the eigenvectors is termed the so-called *score plot*. From a more formal viewpoint, PCA is based on a matrix decomposition method that decomposes ***X*** into two matrices ***V*** and ***U***, the *U*-matrix is also called the *scoring matrix* and contains the above mentioned scores. Contrary to that the ***V***-matrix is termed the *loadings matrix*. In the context of PCA, loadings represent the correlation of each feature of ***X*** with the corresponding PC. Hence a *loading plot* represents a scatter plot of the features in PC space. Loading plots can be used to assess variable importance in an unsupervised manner. However, this can sometimes be highly misleading as suggested in the literature (268–270).

Another way to categorize ML algorithms is to consider the type of the *y*-vector. If the *y*-vector is constituted of continuous values than this is referred to as a ***regression*** problem. In case of a binary *y*-variable this is referred to as a ***classification*** problem.
Accounting for the uncertainties associated with the biological assessment of ABC-transporter ligands, we only make use of binary *y*-variables in this thesis. Hence, the presented models in this contribution are all *classification* models. Based on the fact that supervised learning usually outperforms unsupervised learning, all the different models presented in this thesis can be categorized as: ***supervised, classification models***.

Ensemble Methods vs. Single-learners. A third and final way to discriminate ML algorithms is to divide them according to certain aspects of their implementation. Many contemporary algorithms belong to the class of so-called *ensemble algorithms*. These algorithms are based on the idea to aggregate multiple models into one. Thereby, the aggregated ensemble is supposed to perform better on the learning problem than the single model. Ensemble methods are primarily applied to improve the performance of a model, or to reduce the likelihood of selecting a model that performs poorly on the classification problem. Furthermore, ensemble methods are also commonly used for feature selection tasks. An ensemble is usually generated by iterative sampling of the input matrix $\boldsymbol{X}$ either along its *n*-columns, its *m*-rows or both. In this thesis three ensemble methods are applied: Random Forests, RuleFit, and Boosting. The opposite of ensemble models are the so-called *single learners*. Single learners do not perform any data resembling and therefore base their decision boundary for classification only on one model that embodies all the data. Typical examples of *single learners* are the classical implementations of: *k*NN, SVM, Naive Bayes, Neural Networks.

The main advantages of *ensemble methods* over *single learners* is that they:

- are more powerful with respect to predictive performance and
- tend to be more robust with respect to overfitting.

However, in most cases *ensemble methods* are more difficult to interpret than *single learners*. Most of the results presented in this thesis can be described from a methodological viewpoint as: ***supervised, ensemble classification models***.

Data Preprocessing Methods in Machine Learning

Feature Selection

As explained above the intensive development of many types of different chemical descriptors that encode chemicals in machine readable format has generated large and often over-squared data sets (more features/columns than instances/rows). To extract the most informative and probably most predictive descriptors from a high-dimensional, but often very noisy feature space is the main objective of feature selection algorithms. Feature selection thus, estimates and weights feature relevance. Kohavi and colleagues have categorized features according to their different levels of relevance (271):

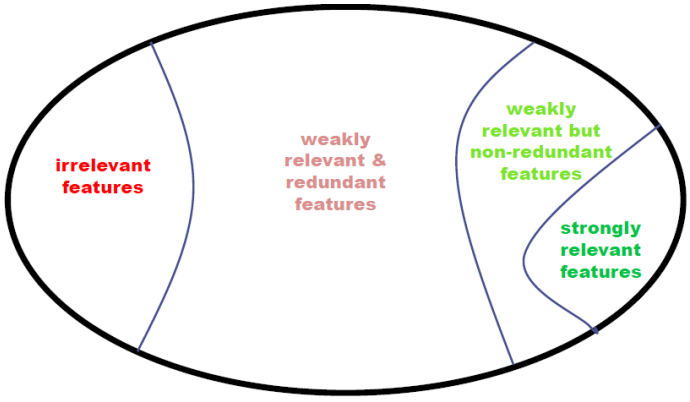


Figure 13: Categorization of different feature types according to their relevance and redundancy. The smaller green regions represent the areas with the most important/interesting features. Adapted and modified from Yu and Liu (272).

Nowadays, a variety of FS algorithms exist. All of them assess the topic of feature relevance in a different context. Some make use of information theory, whereas others assess correlations between descriptors as basis for feature extraction, for instance (273). Furthermore, these algorithms also differ with respect to their algorithmic implementation. Some FS algorithms belong to the class of *filter* methods, whereas others belong to the class of *wrapper* or *embedded methods*. The family of *filter* methods assesses feature relevance using data intrinsic methods, such as class specific distributions or correlations. These methods are usually fast and their computational implementation is often very simple. Contrary to that, *wrapper* methods incorporate the FS pre-pocessing step directly into the machine learning algorithm. These methods are often computationally quite expensive.

In the following the principles of the feature selection problem shall be briefly discussed. The main focus is set to highlight the complexity of the feature selection problem on the following example. A more detailed discussion on this is provided by Isabelle Guyon (274).

Why Univariate Selection sometimes Fails

When one is confronted with the problem to select the most important features out of a set of many to further construct a two-class classification model, an intuitive solution would be to evaluate the features individually on basis of simple univariate statistics. A straightforward solution would be to apply for instance classical hypothesis tests from classical statistics or to use the signal-to-noise ratio (275). As an example, one could apply a t-test on every feature, thereby comparing the class-specific means, and then rank the derived p-values in ascending order or in other words, the smaller the p-value the better the feature. However, such a univariate hypothesis test is highly susceptible to yield unsatisfying results, because it does not consider one of the most crucial aspects of the feature selection problem. *One variable that appears to be useless, when considered on its own, can become very powerful when considered with another variable!* (see schematic A and B below).

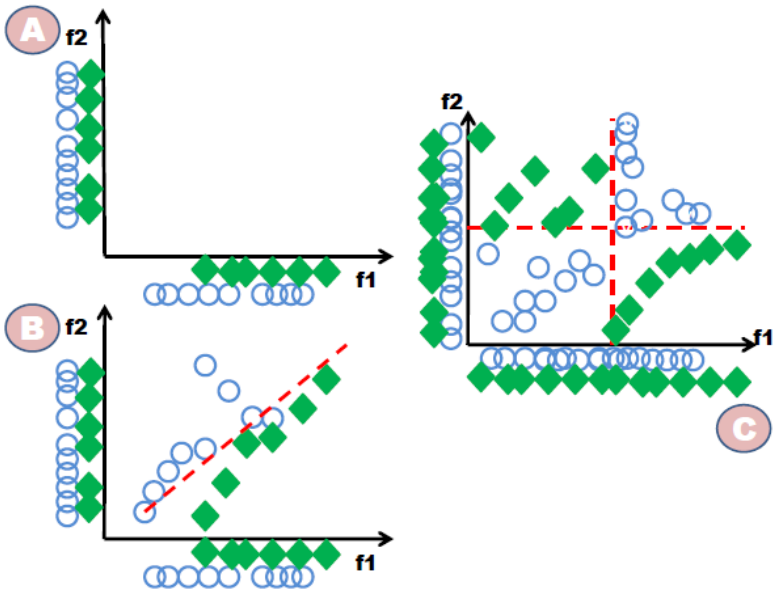


Figure 14: Univariate selection might fail

This example can be escalated further. In many scenarios it can happen that *two* – or in a higher dimensional space even more – *variables appear to be useless when considered individually, but become very powerful when considered together* (see schematic C above). This phenomenon is usually referred to in the literature as *feature interaction* (276). These two examples clearly outline, that feature selection is a highly complex problem and further clarifies why so much research was done on improving these algorithms. The application of feature selection algorithms and their evaluation is a central objective of this thesis.

Selected Feature Selection Algorithms

In the following two selected feature selection algorithms shall be explained in more details. The two algorithms are Information Gain and the Relief method. Both are supervised methods. The reason why these methods have been selected for a more detailed description is because they are relatively simple and easy to understand and embody certain characteristics that can be applied to many different feature selection algorithms and are sometimes also integrated into ML algorithms.

Information Gain. Information Gain (also known as Kullback-Leibler divergence) is a relatively simple FS algorithm that relies upon classical "information theory" and uses the concept of "(information) entropy". It belongs to the class of univariate feature ranking methods. It is a 6-step algorithm and its pseudocode shall be explained in the following on a hypothetical example.

- **Step I:** calculation of the expected information needed to classify a instance in a data set D

	MW	class
Comp.1	280	INACT
Comp.2	245	INACT
Comp.3	467	ACT
Comp.4	890	ACT
Comp.5	678	ACT
Comp.6	782	INACT
Comp.7	437	ACT
Comp.8	101	INACT
Comp.9	205	ACT
Comp.10	691	ACT
Comp.11	285	ACT
Comp.12	367	ACT
Comp.13	491	ACT
Comp.14	1004	INACT

$$Info(D) = -\sum_{i=0}^{n} pi\ log2(pi)$$

$$pi = \frac{Ci}{all\ cases}$$

n=2 → two classes
C1=ACT; C2=INACT

C_i = class membership of class i

$$Info(D) = -\frac{9}{14} * log2\left(\frac{9}{14}\right) - \frac{5}{14} * log2\left(\frac{5}{14}\right) = 0.940$$

Figure 15: Calculation of Information Gain I.

- **Step II:** discretization of continuos values to categorical values

	MW	class
Comp.1	280	INACT
Comp.2	245	INACT
Comp.3	467	ACT
Comp.4	890	ACT
Comp.5	678	ACT
Comp.6	782	INACT
Comp.7	437	ACT
Comp.8	101	INACT
Comp.9	205	ACT
Comp.10	691	ACT
Comp.11	285	ACT
Comp.12	367	ACT
Comp.13	491	ACT
Comp.14	1004	INACT

	MW	MW_disc	class
Comp.1	280	low	INACT
Comp.2	245	low	INACT
Comp.3	467	medium	ACT
Comp.4	890	high	ACT
Comp.5	678	high	ACT
Comp.6	782	high	INACT
Comp.7	437	medium	ACT
Comp.8	101	low	INACT
Comp.9	205	low	ACT
Comp.10	691	high	ACT
Comp.11	285	low	ACT
Comp.12	367	medium	ACT
Comp.13	491	medium	ACT
Comp.14	1004	high	INACT

Figure 16: Calculation of Information Gain II

- **Step III:** calculation of the expected information requirement for an attribute

$$Info(MW) = \frac{5}{14} * \left(-\frac{2}{5} log2 \frac{2}{5} - \frac{3}{5} log2 \frac{3}{5}\right) \quad \text{for low}$$

$$+ \frac{4}{14} * \left(-\frac{4}{4} log2 \frac{4}{4} - \frac{0}{4} log2 \frac{0}{4}\right) \quad \text{for medium}$$

$$+ \frac{5}{14} * \left(-\frac{3}{5} log2 \frac{3}{5} - \frac{2}{5} log2 \frac{2}{5}\right) = 0.694 \quad \text{for high}$$

Figure 17: Calculation of Information Gain III

- **Step IV:** calculation of the gain in information from such a partition

$$Gain(MW) = Info(D) - Info(MW) = 0.940 - 0694 = 0.246$$

Figure 18: Calculation of information Gain IV.

- **Step V:** repeat Steps I to IV for each feature in the data set
- **Step VI:** rank the calculated gain in information (Step IV) for each feature. The feature with the highest value is the most important feature.

One of the critical features of Information Gain is that it is highly efficient in assessing variable importance and is highly useful for identifying the crucial split-points of a feature vector x_j. It is definitely that characteristic, which make it highly appealing and also explains why it is often implemented in decision tree algorithms to calculate decision points. A potential drawback is that it is myopic (i.e. considers only one feature at a time; thus being unable to capture variable interactions). Summarizing, Information Gain describes a simple solution to the feature selection problem and can be used on its own for assessing variable importance, but is also often an integrative part of tree-based ML algorithms, such as decision trees, random Forests or rule-based methods.

Relief Algorithm. Relief is another feature ranking method that also works in a supervised manner. It was first introduced in 1994 by Kira and Rendell (277). Contrary to the algorithm presented above, Relief is a so-called instance or distance-based feature selection algorithm, which is able to capture variable interaction effects (277,278). Another advantage of Relief is that it performs a global search of the feature space as well as a local search, but does not rely on greedy heuristics (i.e. simulated annealing) which probably can get stuck in local minima. The basic concept of Relief is based on a nearest neighbor search. The algorithm randomly picks a learning instance $\boldsymbol{X_i}$ in full feature space and calculates the "nearest hit" (closest instance of the same class) as well as the "nearest miss" (closest hit from the opposing class). Subsequently, the same procedure is carried out in local space (each feature after the other) and then the difference is computed. The Relief algorithm is schematically outlined in the following figure:

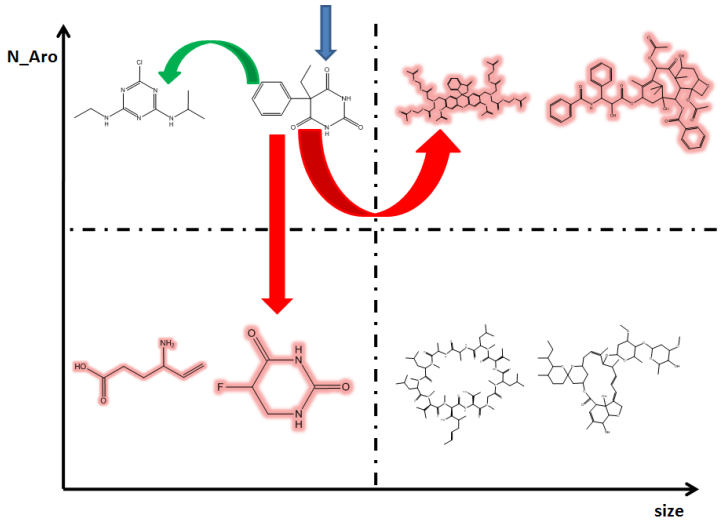


Figure 19: Principle of the Relief FS algorithm: A relevant feature can distinguish between instances of opposing classes although they appear to be closely related. One instance is selected (blue arrow) and the nearest hits and the nearest misses are calculated (green and red arrows respectively). Classes are encoded by shading.

The primary advantage of Relief is definitely that it considers multiple features at a time and that it is still fast, easy to implement and still scales well when input sets are noisy.

Chapter 6: Selected Machine Learning Algorithms

k-Nearest Neighbor Learning

k-nearest neighbor (*k*NN) is a ML algorithm that belongs to the class of the so-called *lazy learners*, which describes its property to classify new unknown instances without prior construction of a decision boundary. In other words learning and classification are done in one single step. It is also called an *instance-based* learner. This algorithm dates back into the 1950s and was first mentioned by Fix and Hodges (279). Because of the fact that this algorithm is computationally quite expensive (especially when subjected to large scale data sets), it was not very popular until increased computing power became available.

*k*NN learning is based on the concept of analogy/similarity. A given test instance is compared to training instances which are similar to it. The training instances are described by their n-dimensional feature vector and each instance represents a point in this n-dimensional feature space. When represented an unknown test instance, the feature space is searched for the *k* training instances that are closest to the unknown test instance. These *k* instances in feature space are the *k*-nearest/closest neighbors of the test instance. A simplified schematic of this algorithm is shown in the figure below. The similarity or "*closeness*" between two or more instances is defined by applying a distance metric. Commonly, for *k*NN learning using numeric attributes the Euclidean distance is used. This distance measure is defined as follows for two points $p = \{x_{11}, x_{12}, \dots, x_{1n}\}$ and $q = \{x_{21}, x_{22}, \dots, x_{2n}\}$ in *n*-dimensional space:

$$dist(p,q) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2}$$

Equation 2: Euclidean distance

In other words, for each feature, the difference between all corresponding values of that feature in *p* and *q* is calculated and squared and finally accumulated. The square root is taken of this total accumulated distance. In order to prevent that attributes with unusual large values (such as molecular weight) outweigh attributes with smaller values (such as number of H-bond donors/acceptors), data is usually scaled via applying either min/max-normalization or z-standardization.

In *k*NN learning the test instance is assigned the most common class label (i.e.: active/inactive or inhbitior/substrate) among its *k*-nearest neighbors.

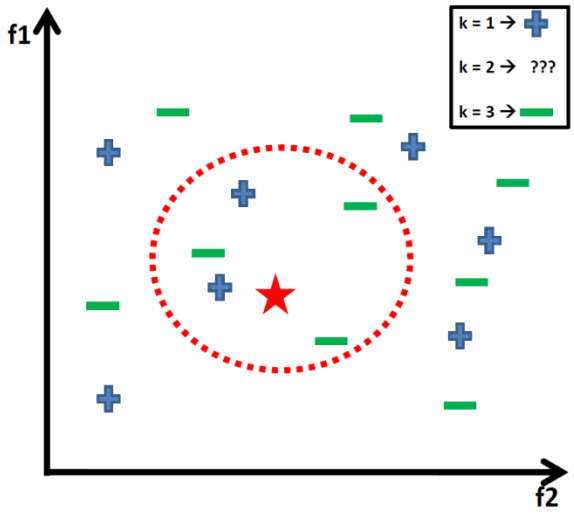


Figure 20: Schematic outline of *k*NN learning. The test instance, with an unknown class label, is projecte into the property space spanned by the instances of the training set. Then the similarity/distance of the test instance to the k-nearest instances of the training set is evaluated and a majority vote for the predicted class label is conducted.

For *k=1* the test instance is assigned that label of the training instance that appears to lie closest to it in feature space. For *k=3* the class which is most common among the three nearest nighbors is assigned to the test instance. From this, it becomes apparent on the one hand that the choice of the spanned feature space and on the other hand the choice of *k* are crucial to the performance of this learning algorithm. The "best" choice of *k* has to be evaluated experimentally and is highly dependent on the data set as this was shown by us (263). However, the most appropriate feature space can be selected by means of feature selection or dimensionality reduction methods.

Although, *k*NN learning seems nowadays to be outdated by more accurate and faster modelling algorithms it is still widely used in the field of drug discovery and ADMET prediction mainly because of its simple and easy-to use implementation (280–287). We also applied *k*NN learning to evaluate different FS algorithms (288). The results of this investigation are outlined in Chapter 12.

Decision Tree Learning

Decision tree (DT) learning, also known as *recursive partitioning* is a class of supervised machine learning algorithms that represent a *flowchart-like* structure, highly similar to a tree. These kinds of algorithms have been mainly developed in the 1980s by J. Ross Quinlan, L. Breiman and J. Friedman (258). A comprehensive overview of work on DTs is given by Murthy (289). The ID3 (Iterative Dichotomiser) (290), the C4.5 (291,292) and the Classification And Regression Tree (CART) algorithm (293), which represent the most popular algorithms from this family, are based on the *divide-and-conquer* principle, which adopts a non-backtracking implementation in which trees are generated in a top-to-bottom, recursive fashion.

The tree-like structure of such implementations is constituted of internal nodes, branches, terminal nodes and a root node. Every *internal node* (or nonleaf node) represents a certain test/question (i.e. is molecular weight higher than 500?) on an attribute of the input data set. Each connecting branch of the tree denotes an outcome of this test (i.e. Yes, molecular weight

is higher than 500). Each terminal node (or leaf node) contains the class label and is used for classification of a new instance that has passed through the whole tree and ended up in a particular leaf node. The topmost node of a DT is the so called root node. The feature that divides the training instances best is assigned to represent this top root node. A variety of different methods have so far been implemented to evaluate and select this feature and to define the optimal split point for class separation. Among the most popular of these methods is the Information gain, the gini index (293) and the ReliefF algorithm (278). A general architecture of a DT is given in the schematic below. DTs are definitely among the most popular ML algorithms ever invented.

The main reasons for their popularity are probably that they:

- do not require any parameter setting,
- can handle high dimensional data well,
- are able to select the most relevant features
- and are easy to interpret.

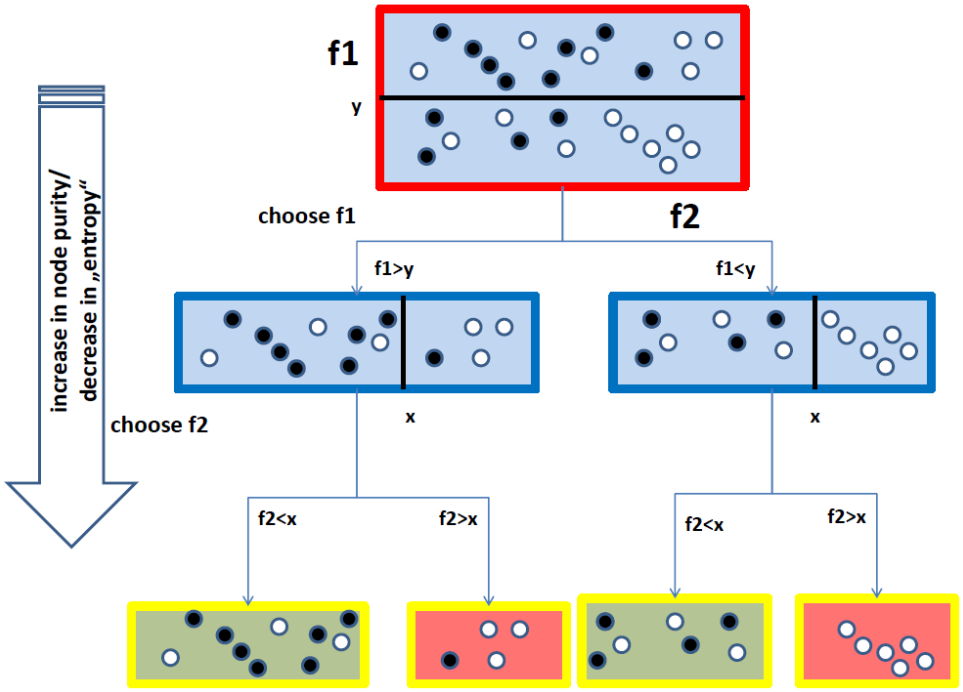


Figure 21: Schematic illustration of a decision tree algorithm

Despite all these advantages DTs harbor also several disadvantages. They are highly sensitive to overfitting (i.e. perfectly explain the training data, but perform weak on new test sets), which mainly relies on the fact that even small perturbations regarding the attributes can highly affect classification accuracy. Further, they are relatively intolerant to redundant or highly interdependent attributes.

Although DTs are not explicitly used in this thesis, they constitute the algorithmic basis of the following two algorithms explained.

Random Forests Learning

Random Forests (RF) is another classification algorithm that belongs to the family of *ensemble learning algorithms*. Such ensemble learning methods combine several single learners to a final model. This procedure aims to yield models that have more predictive power than the single models. Formally, RF can be seen as a collection of individual decision trees. Hence, the name Random Forests. The algorithm has first been presented in 2001 by Leo Breiman (294). The algorithm, which is employed by RF is graphically outlined below. Briefly, RF constructs each of its trees in the following manner: First, a bootstrap sample (i.e. sampling with replacement) of the instances is drawn. Similarly, a prespecified number of features (by convention this is called m_{try}) is randomly selected. This reduced data set is consecutively used as input to construct the first decision tree, which is fully grown. In parallel the rest of the instances which is not considered for the construction of this particular tree is predicted to assess the prediction error of this tree. This procedure is repeated *n* times until *n* decision trees are built.

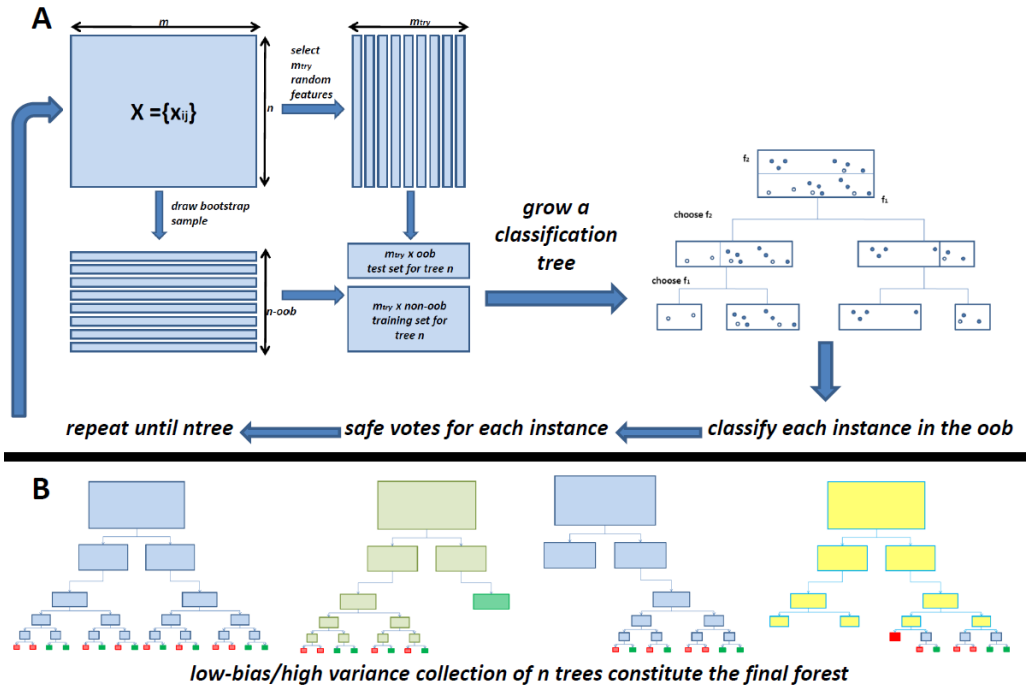


Figure 22: Schematic illustration of Random Forests learning

A new data set can be predicted from the collection of these decision trees as every instance is run down the tree and is assigned the class the terminal node of this tree contains. This procedure is then repreated over all trees in the Forests and the final class assignment is done via a majority vote. This majority vote is by definition a binary response (e.g.: 0 for inactives, 1 for actives). However, RF also provides a class probability estimate which is simply the fraction of votes for a particular class. For instance, if a compound is assigned the positive class by 400 trees and the opposing class by 100 trees in a RF that is constituted out of 500 trees, it will be assigned a probability of 0.80 for the positive class.

Based on this algorithmic features, RF comprises a lot of advantages compared to other ML methods (262,265,295).

RF modeling:

- is highly accurate
- is highly efficient for large-scale datasets
- can deal with oversquared input matrices (e.g.: more descriptors than compounds)
- is rather fast, because each tree is constructed only on a subsample of the input space
- requires almost no tuning of hyperparameters
- can efficiently handle unbalanced class distributions
- can be used for classification and regression problems
- provides estimates of variable importance
- can be used in unsupervised mode, using its intrinsic proximity measure (RF clustering)
- provides appealing features of model visualization
- theoretically, no CV is necessary

Although all these properties of the RF algorithm are very appealing, some disadvantages have also been observed. A study by Segal showed that RF might tend to overfit for some datasets (296). However, it needs to be stated that the data on this issue is controversial (294). The study by Segal only evaluated regression problems. Furthermore it was observed that RF is inferior to many other - and probably more simple - algorithms when challenged with multi-class problems. Additionally, it seems not to be highly efficient for categorical variables with different levels (297). Given the high popularity of RF many open-source implementations have emerged during the past decade. In this thesis we made use of the RF implementation in *R* in the *randomForests*-package obtained from CRAN (298). RF is used in Chapter 13, 14 and 15 as modelling algorithm in this thesis.

Rule-based Learning

In the two sections above, decision trees and their extension random Forests have been described. It has been explained that their flowchart-like tree structure origins from certain tests (questions) on an attribute. Hence, the classification model generated by a decision tree is an accumulation of such tests (e.g.: has the molecule more than five hydrogen-bond acceptors?). From the viewpoint of rule-based learning, such tests also represent the rules of such a model. In other words, the flowchart-like architecture of a decision tree can also be considered as a set of different rules that characterize a particular data set and can therefore be connected to establish a decision boundary that can be utilized for classification. These rules are usually connected via boolean operators, such as logical AND (&&) or logical OR (||). A schematic overview of this implementation is given in the following figure:

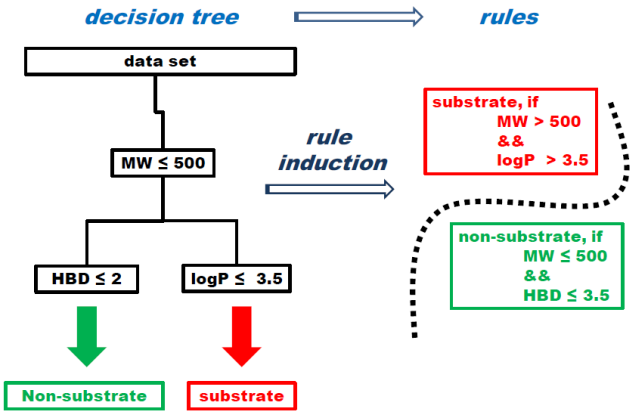


Figure 23: Rule induction from a decision tree. A decision tree is converted into a set of connected, logical rules. The black dotted line denotes the non-linear decision boundary of the rule-based model. MW=molecular weight; HBD = hydrogen-bond donor; logP=log of the octanol/water partition coefficient (≈ lipophilicity).

Therefore, it can be generalized that decision trees can be converted into a set of logical rules. This process is referred to as *rule induction*. It should be noted that decision trees often harbour many of these rules, and therefore the figure above is highly simplified. Furthermore, there are also other ways to conduct rule induction, such as association rules (for more details the interested reader is referred to the textbook by Frank and Witten and also to chapter 14 of the textbook of Hastie and Tibshirani (262,265)). As decision trees belong to the class of non-linear algorithms, it is obvious that rules also classify instances in a non-linear fashion. However, in order to construct a classification algorithm, which utilizes rules, these rules must often be implemented into classical linear decision functions. In this process, weights w_i are assigned to the individual rules. These weights can readily be used for interpretation purposes of the rule-based model. Both, the sign and the magnitude of b_i dictate the interpretation of a given $Rule_i$. The sign indicates for which class a particular class "votes", whereas the magnitude of w_i indicates the importance of the rule for the model. This is illustrated below. Furthermore, many rule-based algorithms also calculate other numeric values, which are also useful for interpretation of such a model. These values are called *rule support (s)*. The support value *s* indicates reports the co-occurrence of different items (e.g.: from the example above: MW && logP) of a rule. The higher the *s* value, the more reliable a given rule.

$$Rule_1, Rule_2, \dots, Rule_n \longrightarrow f(x) = \sum_{i=1}^{n} w_i \; x \; Rule_i + \varepsilon$$

$$substrate\ if, \qquad f(x) \leq 0$$

$$non - substrate\ if, f(x) > 0$$

Figure 24: Rule-based Classification models. Individual rules are integrated into a linear decision function. Individual rules are weighted and summed up. This finally returns a numeric output which can then be used for classification.

In summary, rule-based models are very popular classification models, since they can be easily interpreted and in case the rules are based on simple concepts medicinal chemist are

familiar with (such as MW and logP), they can also be readily implemented into real-world applications. However, the main disadvantage of rule-based classification is that in analogy to a simple decision tree, they are susceptible to small perturbations in the inputs and therefore tend to overfit the data which jeopardizes performance estimation. In order to overcome this problem, ensemble rule-based algorithms have been invented. In this thesis, such an ensemble algorithm, *RuleFit* by Friedman and Popescu is applied to classify ABCB1 substrates/non-substrates (299).

Support Vector Machines

Support vector machines (SVM) constitute another type of ML algorithms that are based on entirely different definitions than the algorithms already presented. SVMS were first presented by Cortes and Vapnik in 1995 in an attempt to implement the Structural Risk Minimization principle of statistical learning theory (300). The basic concept of SVMs is to separate data points, which belong to different classes by constructing a separating line in the input space spanned by the data set under investigation. However, this problem is not trivial, since many different class-separating lines can be computed in the input space (for illustration see a simplified problem in 2D space shown below in A). In order to solve this problem, the algorithm searches for the optimal separating hyperplane, which maximizes the distance to the nearest data points from opposing classes (see figure B below). Hence, SVMs are also termed *maximum-margin* classifiers. These instances, are the only ones that are used to compute that hyperplane and therefore these instances are called *"support vectors"*.

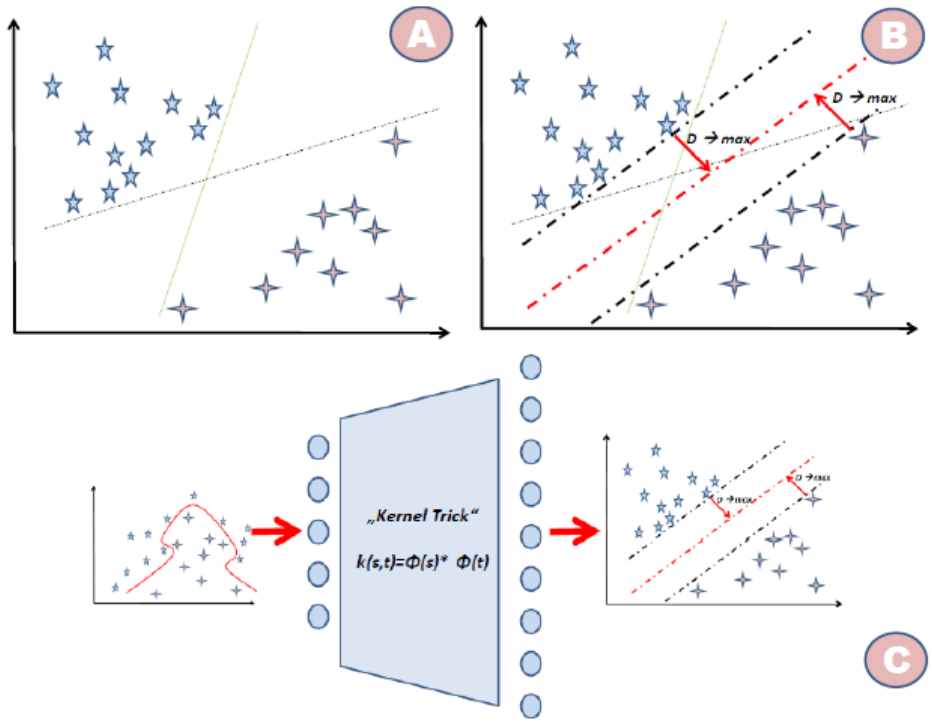


Figure 25: Principles of SVM classification illustrated using an hypothetical 2D example. A) Problem of finding an optimal class-separating decision line. B) maximum-margin decision line. C) "Kernel trick": subjecting the linearly non-separable data into higher dimensional feature space to find a linear maximum-margin hyperplane in that transdormed space.

Another important feature of SVM modelling is that it is especially useful for data sets for which linear class separation is not possible. For such non-linear separable data sets, SVMs perform a data transformation, which projects the input space into the so-called "transformed feature space" (301). This transformation is caused by a kernel-function. The main motivation of the application of such a kernel-function is that probably a linear separating hyperplane is found in that transformed feature space that cannot be constructed in the original input space (see figure C above). This is called the "kernel-trick" (302). It is important to mention that different kernels have been developed in the last decade. Commonly used kernels are the polynomial-, the radial basis function-, the exponential-, and the linear kernel. A detailed overview of Kernel methods with a special focus of their application in chemistry is provided by Ivanciuc (303). In summary, SVMs are very popular learning machines in the field of drug discovery since they are very efficient to capture nonlinear relationships between the inputs and their class label. However, the main disadvantage is that SVMs require a lot of parameter tuning and furthermore, the resulting model is usually difficult to interpret. In this thesis SVMs are used in Chapter 15 for the classification of MDR-selective molecules.

Chapter 7: Validation of Machine Learning Models

Types of Model Validation

The validation of a ML model is a key step in the ML development process. Considering, that ML models serve the primary utility to be used to reliable predict new chemicals, it becomes apparent that the estimation of their performance is crucial. Various validation criteria have been proposed and different types of validation methods exist. These methods can be grouped into two categories:

- **Internal validation**
- **External validation**

Internal validation methods give answer to the question: "*Can we believe, what the model tells us?*". These methods describe the so-called *"efficacy"* of a model. Contrary to that, external validation methods give answer to the question: *"Can we use the model to reliable predict new instances?"*. External validation describes the *"generalizability"* or in other words the *"effectiveness"* of a model. It is important to note, that good internal predictivity is a prerequisite for external preictivity.

The main difference between internal and external validation techniques is that the former make use of all the available data, whereas the latter require the definition of a hold-out prediction/test set that does not participate in model generation. Ideally, such a test set should contain new data that are retrieved from external sources. However, in real-world practice pharmacological data is often scarce and therefore the input sets are often split *a priori* into a training set (which is used in model construction) and a "supposedly unknown" test set (which is used for validation only). The main advantage of the external validation method is that it is very fast and computationally inexpensive. Furthermore, it is probably the best strategy to estimate putative model overfitting (= model performs good in training, but performs poorly on application tasks). However, the main drawback is that only a reduced amount of all the available information is used for model construction. Typically, a test set contains 20% of the available information, whereas the training data consist of 80% of the data. Furthermore, the heuristic applied to generate the test set can also influence the final performance. It can be summaried that a "good" external test set shall be as representative as possible of the whole input space (see the following schematic for illustration). Several strategies to define such external test sets have been proposed (see the work of Golbraikh et al. (304,305)), such as random splitting or splitting on basis of input data diversity (306). Internal validation techniques make use of the whole body of data and estimate the predictivity of a ML model using different iteration techniques. At each iteration a certain amount of instances is hold-out for model generation and used for estimation of the model`s performance to predict them. Such iterative, internal validation techniques are termed crossvalidation. Depending on the number of molecules that are put aside at each iteration, the following different techniques can be distinguished: Leave-one-out (LOO) crossvalidation, Leave-multiple-out (LMO) crossvalidation. The latter strategies are also called *k*-fold crossvalidation techniques and *k* describes the number of iterations.

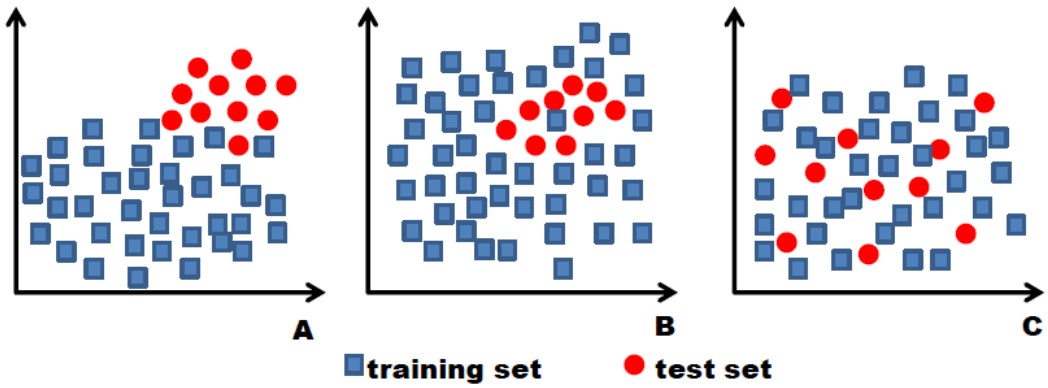


Figure 26: Selection of training and test sets. The method determines whether "good" or "bad" test sets are derived.

Thus, a 10-fold crossvalidation embodies 10 iterations and at each iteration 10% of the input instances are put aside, whereas in 5-fold crossvalidation 5 iterations and 20% of the input molecules are used at each iteration. A general pseudocode for a simple LOO-cross validation scheme runs as follows:

- *for i=1 to k*
  - *let (x_i,y_i) be the i^{th} data entry*
  - *temporarily remove (x_i,y_i) from the input data*
  - *train a model on the remaining i-1 compounds*
  - *predict (xi,yi)*
- *report the mean performance*

A comparative overview on the different properties of internal and external validation methods is summarised in the following table:

Table 3: Properties and Features of different Validation Techniques

	Internal validation	**External validation**
Purpose	efficacy	effectiveness
Aim	to demonstrate robustness and stability of the model	to demonstrate generalizability of the model
Principle	iterative subsampling of all data	definition of a test set
Advantages	• makes use of all data	• computational inexpensive • estimates overfitting
Disadvantages	• computational expensive	• generates reduced data sets • dependent on the selection algorithm

A comprehensive overview of contemporary validation criteria to assess the "real external predictivity" of QSAR models (but with a focus on regression) is given in two comprehensive articles published by Chirico and Gramatica. (307,308).

Measures of Classification Performance

The output of a supervised classification model is usually either a voted label for a particular class for each instance or a probability vector, which denotes an estimated probability for each compound belonging to a particular class. In the latter case, it is necessary to define a probability threshold to allocate the predicted compound to one of the classes. In many studies, a threshold of >0.5 is applied to discretize the output probability vector of the model into a binary class label. In order to calculate measures of classification performance the predicted class labels are compared to the actual (measured or observed) class labels and subsequently true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) are arranged in a *confusion matrix*. Confusion matrices have the following properties:

- they represent square matrices
- their dimensionality equals the number of classes (e.g. two classes → 2x2; three classes → 3x3)
- correct predictions are found at the diagonal of the matrix, whereas misclassifications are located at off-diagonal elements
- the columns summarize the performance of the particular classes.

Fom the derived confusion matrix different measures of model performance can easily be calculated (see schematic below). These measures are thoroughly explained in Baldi et al. (309).

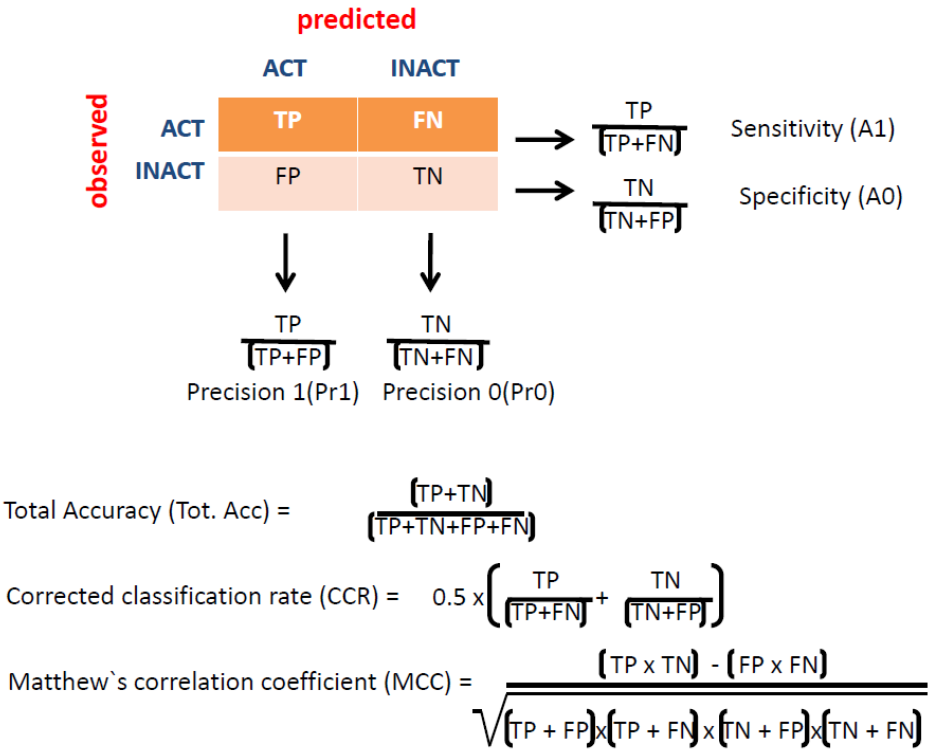


Figure 27: Different measures of classification performance

It is important to note that these measures can be categorised roughly into two groups:

- **class-specific measures**
  - sensitivity, specificity
  - precision
- **measures that consider both classes**
  - total accuracy
  - corrected classification rate (CCR)
  - Matthew`s correlation coefficient (MCC)

Most of these different measures return values in the range [0,+1]. The only exception is MCC, which can take values in the range [-1,+1]. All these values can be interpreted in the same way: the closer the calculated value to one, the better the performance of the model. It is important to mention that none of these measures alone can reliably describe the performance of the ML.

The Applicability Domain of Machine Learning Models

As outlined above ML models are usually constructed to serve two primary purposes. At first, they shall provide information on the chemical, structural or physicochemical patterns that are associated with a given property. Such information is usually inferred by means of model interpretation, which can be done via assessing the different variable importances, for instance. In order to reliably interpret a ML model, high internal validitiy of the model is usually sufficient. Another aspect of a ML model is to use it with the aim to predict certain properties of new compounds, with unknown liability to the property of interest. In order to

provide an estimate on the capability of a model to predict such new unknown compounds, external validation is usually a prerequisite. However, many models have been shown to provide very unsatisfying results, as soon as they have been used for the prediction of new molecules (310). This can be referred to as an "efficacy-effectiveness gap"; i.e. models perform well during the implementation phase, but fail when they are used in practical, real-life situations. This phenomenon has urged the community to define the so-called applicability domain (AD) of a model. Roughly, the AD is an additional measure or maybe several measures, which describe a certain degree of confidence for the models prediction (see the figure below for a simplified visualization). The estimation of the AD of a ML model belongs to the model post-processing techniques and they are applied with the primary objective to decide whether a compound with unknown activity/property will be predicted well by the model (311). The OECD guideline for the "Principles for the Validation of QSARs" (312) defines the applicability domain of a ML model according to Netzeva et al. as follows: *"The applicability domain of a (Q)SAR model is the response and chemical structure space in which the model makes predictions with a given reliability."*(313).
This definition was first described in 2005 and since then a lot of research has been done to derive at reliable measures for the determination of the AD of a ML model (284,314–317).

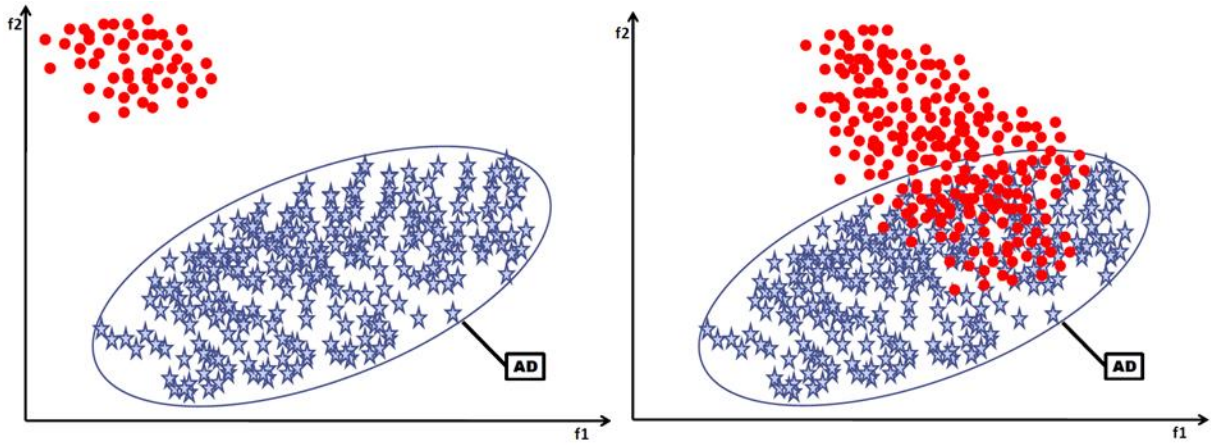


Figure 28: The concept of AD: Out of intuition a model trained on a given set of compounds is unlikely to reliably predict the "whole chemical universe". Left panel: red molecules are outside the AD; right panel = overlap between model AD and molecules outside the domain. blue stars = molecules used for modelling; red circles = new molecules with unknown activity/property.

According to Jaworska and Tetko the different AD measures can be categorized into the following groups (318,319):

- **range-based methods –** represent the most simple form of AD estimation methods. These methods simply approximate the convex hull of the training data by considering the ranges (from minimum to maximum) of the individual features. The major drawback of this methods are that they expect the data to be normally distributed, otherwise there approximation of the convex hull would enclose considerable empty space.
- **distance-based methods –** these methods calculate the distance from each point in the training set to a particular point in the test set. The three commonly used distance measures in ML include: the Euclidean, Mahalanobis distance and City block

distances. All these methods require a predefined definition for a threshold by the user. This threshold should be evaluated empirically and might differ from data set to data set. Furthermore the different distance measures make different assumptions about the data. Euclidean and Mahalanobis distances require input data to be normally distributed and expect orthognality between features. The City-block distance is especially useful for discrete descriptors, such as fingerprints.
Additionally, for regression problems the Hotelling T2 and *leverage* are often used to estimate the AD of QSAR models. The *leverage* refers to the diagonal elements of the hat matrix *H* and usually a value >3 is used as a cutoff for judging predictions as unreliable.

- **geometric methods –** aim to approximate the closest convex hull that encloses all training data points. These methods are usually computationally quite expensive, especially in high dimensional feature spaces. A potential drawback of these methods is that they cannot identify potential interior empty spaces.
- **probability density distribution methods –** these methods calculate a probability for each test molecule of belonging to the training set. In general, two differnet methods are commonly applied: *parametric methods*, which make use of Gaussian or Poisson distributions, and *non-parametric methods*, which estimate the probability density from the input data. These kind of methods are the only ones, which are able to discover dense or sparse regions of the training set. However, these methods become computationally very costly as data dimensionality increases.

A particular drawback of the relatively "young" AD field is that most measures presented so far are only applicable for regression problems and only a minority of methods can be used for classification problems. Especially, more sophisticated methods that also make use of the predicted values are only suitable for regression models. For instance, it has been proposed that using the standard deviation calculated from the predicted values of different models can be used to estimate the AD; the lower the standard deviation, the higher the reliability that one can trust the predicted value. Unfortunately, such appealing approaches are not suitable for classification models which often report only a class vote. Additionally, such an approach expects a given variance between the different models. It also needs to be considered, that different AD measures will assess the AD in different ways, and therefore results might be substantially different from method to method. Therefore it has been recently proposed that multiple AD measures shall be combined into one measure. Furthermore, it needs to be decided how to handle those molecules which are predicted to be out of the domain.

Nevertheless, the estimation of a models AD has gained a lot of poularity in the last few years and the incorporation of prediction reliability in classification QSAR models might represent a clear improvement for the ML field.

The estimation of a models AD is also a particular objective of this thesis.

Chapter 8: Network-like Similarity Graphs to Explore Structure-Activity Relationship and Structure-Selectivity Relationship Data

Another and conceptually different way to analyse drug discovery data (compared to the ML methods explained in the previous sections) is to make use of molecular network representations. In general, networks of structure-activity relationship (SAR) data can be either used to graphically represent drug-target networks or to visualize so called "SAR landscapes". The first application is widely used in the field of chemogenomics to analyse cross-pharmacology relationships.

Here the center of attention is set to the analysis of SAR landscapes. These landscapes characterize sets of molecules with respect to changes in chemical structure and corresponding changes in biological/pharmacological effects.

Characterization of SAR Types via Activity Landscapes

According to the chemical similarity principle, which states: "*molecules that are similar with respect to their chemical structure, also show similar biological activities*", one would expect that continuous chemical changes in a lead series result in continuous changes in the activity of these molecules (320). However, in an attempt to systematically analyze principal SAR features, Peltason uncovered in a molecular similarity analysis that most SAR types are very heterogenuos in nature; i.e. even small changes in chemical structure, can massively influence biological activity (321). This finding gave rise to the concept of SAR landscapes, which conceptualizes topological maps of potency distributions in chemical space (322). In these topology maps, the potency of a molecule is added as a third dimension to a 2D projection of the input space (323). According to the topological appearance of a compound series several different SAR types can be defined (see also Fig. below):

- **"rolling hill-like" SARs:** This type characterizes compounds with similar potency, but different chemical structures. It is also called a "**continuous**" SAR.
- **"rugged" SARs:** Here compounds having significantly different potency, exhibit quite similar chemical structures. This type is also referred to as "**discontinuos**" SAR.
- **"activity cliffs":** The last SAR type shows a canyon-like topological appearance that is formed by structurally highly similar molecules that show high potency differences. Often such cliffs are formed by pairs of structural analogues.

Data sets that show different SAR types are termed **heterogeneous SARs** and they can be considered very useful for medicinal chemists in the context of lead-identification and lead-optimization, since they enable the identification of structurally diverse active candidates (in rolling hill areas) and also offer the potential for optimization (if these compounds are close to cliff regions). Topological SAR characteristics of different data sets and their visualization using molecular network representations allows the systematic analysis of large scale data sets across many targets simultaneously and also allows the characterization of compound selectivity in a data-oriented fashion (324). Hence, the concept of SAR landscapes also applies to structure-selectivity relationships (SSRs).

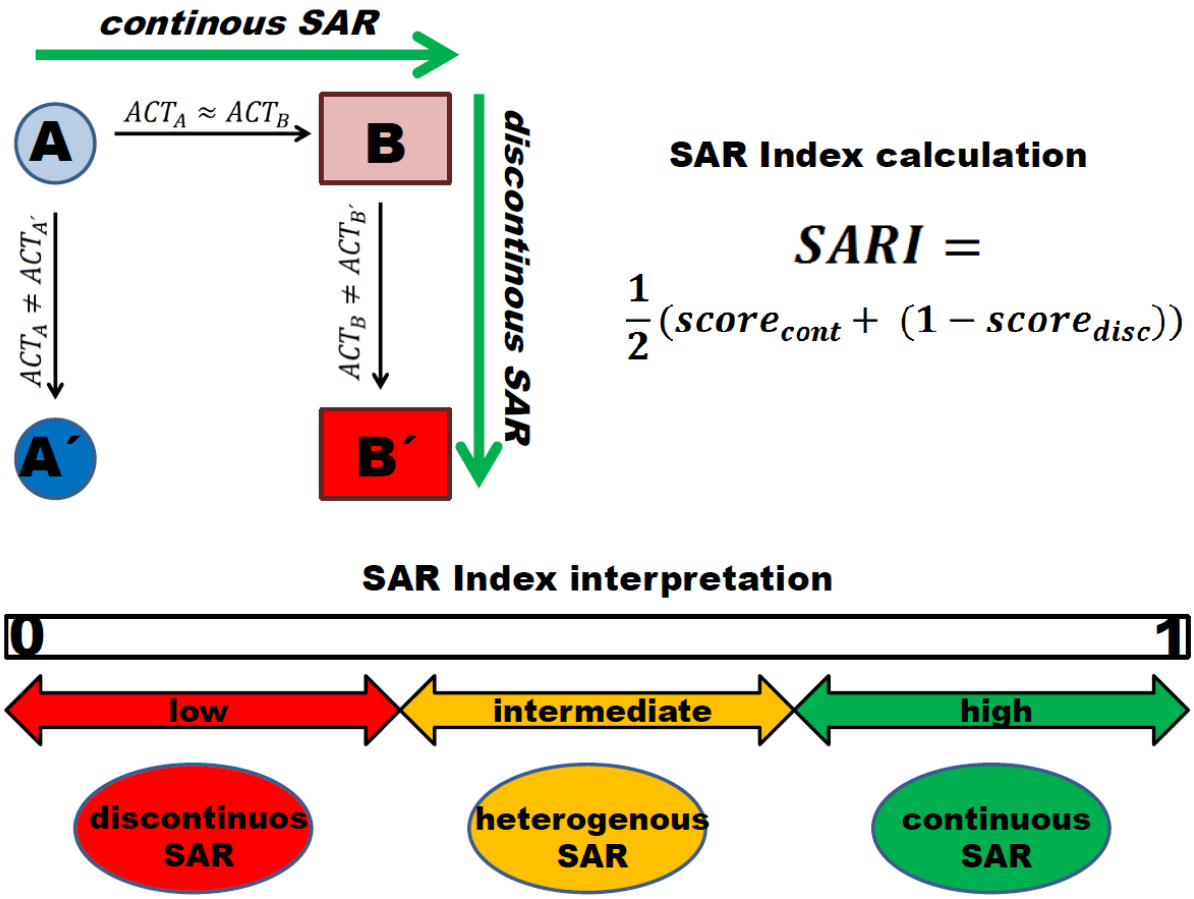


Figure 29: SAR types and their categorization. The upper left panel describes the relationship between continuous and discontinuous SAR landscapes. Two structurally different molecules A and B (light blue circle and light red square) have similar activity values, whereas closely related analogues of A and B show (dark blue circle and red square) remarkable potency differences. Hence, these four molecules give rise to different SAR types. The upper right panel shows the formulation of the SARI, which is constituted of a continuous and a discontinuous function. Additionally, the interpretation and categorization into the different SAR types is graphically shown.

Quantification of SAR Types via SAR Indices

In order to provide quantitative insights into the different SAR landscapes described above, the group of Jürgen Bajorath (in collaboration with R. Guha) extensively elaborated a set of numeric functions that aim to complement these qualitative representations (325,326). These formulations are termed SAR indices (SARIs) and can be calculated for whole data sets (global SARI) or individual molecules of particular data sets (local SARI). Noteworthy, the calculation of SARIs combines molecular similarity and biological activity.

Global SAR Indices

For whole data sets with available 2D chemical representation of molecules and their associated continuous biological activity (determined either as (p)IC_{50}, (p)K_i or as selectivity index), the global SARI captures aspects of continuous SAR regions as well as aspects of discontinuous SAR regions in one numeric value that ranges from 0 to 1. Hence, the global SARI score consists of two individual scores; the global continuity score ($score_{cont}$) and the global discontinuity score ($score_{disc}$). The formula of the SARI is shown in the figure above. Using this formula Peltason and Bajorath were able to categorize the following SAR types on basis of the SARI:

- **high** SARI score as a result of **high** $score_{cont}$ and **low** $score_{disc}$
  - indicate **continuous** SAR types
- **low** SARI score as a result of **low** $score_{cont}$ and **high** $score_{disc}$
  - indicate **discontinuous** SAR types
- **intermediate** SARI score as a result of **high** $score_{cont}$ and **high** $score_{disc}$
  - **coexistence** of continuous and discontinuous SAR types
    - **heterogeneous** SAR type
- **intermediate** SARI score as a result of **low** $score_{cont}$ and **low** $score_{disc}$
  - a continuous SAR within the limits of a **structural constraint**
    - **heterogeneous-constraint** SAR type

The last two categories are both characterized by intermediate SARI scores and describe heterogeneous SARs that can be distinguished according to the magnitude of their $score_{cont}$ and $score_{disc}$ values (325).

It is important to note that the $score_{cont}$ measures the potency –weighted structural diversity within a given compound set, whereas the $score_{disc}$ captures the average potency difference among similar compound pairs (325).

According to this, the SARI and its different subscores provide an easy-to-calculate and useful means to quantitatively describe the nature of SARs for whole data sets. Furthermore, the SARI scoring scheme (when calculated for different data sets) can be used as a highly sensitive measure to estimate the inherent SAR characteristics within different data sets.

Local SAR Indices

The measures presented above provide information on the characteristics of whole data sets and can be used to identify sets that contain actvitiy cliffs. However, if one wants to identify those compounds within a data set that are responsible for these cliffs, a local variant of $score_{disc}$ must be calculated. Such a local discontinuity score has been proposed by Wawer in 2008 and it considers the individual compound contribution to the overall SAR landscape. It considers only molecules that are within a predefined similarity threshold (e.g. Tanimoto coefficient > 0.65). This local $score_{disc}$ is also normalized so that it takes values in the range from 0 to 1. Higher values indicate the activity/selectivity cliff markers contained in a data set.

Visualization of SAR Types using Network-like Similarity Graphs

In *in-silico* drug discovery, network representations have gained wide acceptance to model ligand-target relationships or for quantifying the relationships between different classes of drug molecules (327–330). In order to graphically depict the relationship between compound similarity and potency or selectivity, Network-like similiarity graphs (NSGs) can be employed. In principle these network graphs consist of nodes and edges. The nodes represent the compounds and if two or more compounds exceed a pre-defined similarity threshold than they are connected via an edge.

General Methodology

In order to connect the nodes in a NSG, pairwise compound similarities are calculated using a similarity measure (e.g. Tanimoto coefficient - Tc) from a fingerprint representation of the molecules. Subsequently, this square similarity matrix is transformed into another square binary adjacency matrix. This adjacency matrix contains "1" if the corresponding molecule pair exceeds the similarity threshold and contains "0" if the pair shows a smaller similarity measure than the threshold. This matrix is used to connect the nodes and hence places and connects the molecules of a data set in a 2D space. In order to introduce the activity (pIC_{50} or pK_i) of each compound as a "third" dimension of information, nodes are color-coded using a continuous spectrum (e.g. from green (low potency/selectivity) via yellow to red (high potency/selectivity)). It is important to note, that such NSG representations require a continuous activity value (regression) and cannot be used for binary activity data sets (classification). Furthermore, the size of the node encodes the local compound discontinuity score described above. The larger the node, the higher the local $score_{disc}$. In other words, the larger the diameter of a node, the more likely that this molecule represents an activity/selectivity cliff.

Fruchterman-Reingold Algorithm

The topology of NSGs is determined by pairwise similarity relationships and the distribution of activity values within a data set. The overall layout of such a network is determined by applying the Fruchterman-Reingold algorithm. This algorithm has been proposed in 1991 and belongs to the class of force-directed layout algorithms. In this algorithm network nodes can be considered as electrically charged particles that can repulse each other depending on the distance between these nodes (note the analogy to Coulomb`s law). Additionally, the edges are considered (in analogy to Hooke`s law) as springs, that push connected nodes closer together. The underlying idea is to minimize the energy of the system by iteratively moving the nodes until an equilibrium state is achieved. An important feature of this algorithm is, that it does not scale the distances between nodes by similarity values and therefore, the length of the edge is only different between different nodes because of representation purposes. The resulting network topology generated by this algorithm guarantees that similar nodes (compounds) are placed in close vicinity to each other while at the same time dissimilar nodes are localized far away from each other. Hence, this algorithm is very useful for visualizing large undirected networks in a simple, intuitive and also interactive manner.

Interpretation of NSGs

NSGs are highly suitable to convey the following different levels of principle SAR information:

- **pairwise similarity relationships**
  Nodes that represent the molecules in a data set are connected via an edge if their pairwise similarity exceeds a threshold.
- **potency distribution**

Color-coding of the nodes is used to highlight the spectrum of different potency values within a data set. In selectivity NSGs the node color describes the calculated selectivity index.

- **SAR discontinuity and activity/selectivity cliffs**
  The node size visually depicts the contribution of an individual molecule to the overall discontinuity score of the whole data set.

The following schematic illustrates these different levels of information that can be revealed from an NSG.

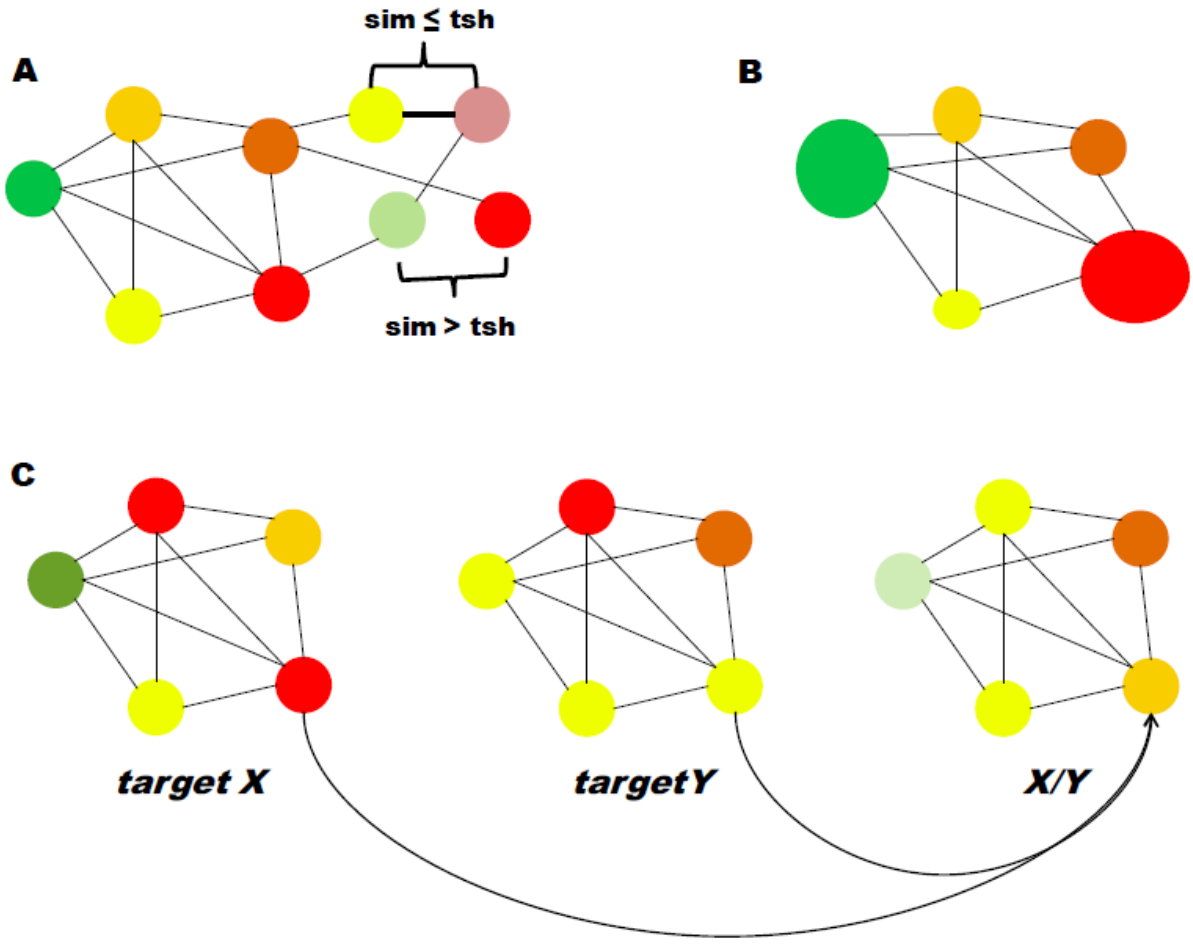


Figure 30: Depiction of NSG principle information levels. Nodes represent represent molecules and edges connect similar nodes (molecules). A) Nodes are connected if their pairwise similarity exceeds a predefined similarity threshold. The color-coding depicts the biological activity of each node. Note, that the length of the edge does not convey information on the actual similarity value, but only reports a binary statement if the similarity is below the threshold or not. B) Identification of activity cliff markers by considering node size. Compounds showing larger node diameters are likely to represent activity cliff markers, or at least a region of local discontinuous SAR landscapes. C) The derivation of a selectivity network by merging the NSGs of compounds with activity values for two different targets.

Summary

A particular advantage of NSG representations for SAR data is that they convey multilevel SAR information in a very intuitive manner and they can easily be "updated" when more active molecules become available. Despite these very appealing properties of NSGs it must also be mentioned that they cannot be used for property predictions in a prospective context, like ML algorithms described in the chapters above, but are only useful for

retrospective data analysis. Nevertheless, they can be used for systematic analysis of global and local SAR regions and can easily identify critical SAR phenotypes that require optimization in lead projects or that can be exploited for scaffold hopping. Such NSG representations are used in this thesis to explore the SAR landscape of a MDR-selective cytostatic compound series (Chapter 16).

References

1. Kessel D, Botterill V, Wodinsky I. Uptake and retention of daunomycin by mouse leukemic cells as factors in drug response. Cancer Res. 1968 Mai;28(5):938–41.

2. Juliano RL, Ling V. A surface glycoprotein modulating drug permeability in Chinese hamster ovary cell mutants. Biochim. Biophys. Acta. 1976 Nov 11;455(1):152–62.

3. Tsuruo T, Iida H, Tsukagoshi S, Sakurai Y. Overcoming of vincristine resistance in P388 leukemia in vivo and in vitro through enhanced cytotoxicity of vincristine and vinblastine by verapamil. Cancer Res. 1981 Mai;41(5):1967–72.

4. Bell DR, Gerlach JH, Kartner N, Buick RN, Ling V. Detection of P-glycoprotein in ovarian cancer: a molecular marker associated with multidrug resistance. J. Clin. Oncol. 1985 März;3(3):311–5.

5. Germann UA, Willingham MC, Pastan I, Gottesman MM. Expression of the human multidrug transporter in insect cells by a recombinant baculovirus. Biochemistry. 1990 März 6;29(9):2295–303.

6. Cole SP, Bhardwaj G, Gerlach JH, Mackie JE, Grant CE, Almquist KC, u. a. Overexpression of a transporter gene in a multidrug-resistant human lung cancer cell line. Science. 1992 Dez 4;258(5088):1650–4.

7. Chaudhary PM, Roninson IB. Expression and activity of P-glycoprotein, a multidrug efflux pump, in human hematopoietic stem cells. Cell. 1991 Juli 12;66(1):85–94.

8. Schinkel AH, Smit JJ, van Tellingen O, Beijnen JH, Wagenaar E, van Deemter L, u. a. Disruption of the mouse mdr1a P-glycoprotein gene leads to a deficiency in the blood-brain barrier and to increased sensitivity to drugs. Cell. 1994 Mai 20;77(4):491–502.

9. Cole SP, Sparks KE, Fraser K, Loe DW, Grant CE, Wilson GM, u. a. Pharmacological characterization of multidrug resistant MRP-transfected human tumor cells. Cancer Res. 1994 Nov 15;54(22):5902–10.

10. Dey S, Ramachandra M, Pastan I, Gottesman MM, Ambudkar SV. Evidence for two nonidentical drug-interaction sites in the human P-glycoprotein. Proc. Natl. Acad. Sci. U.S.A. 1997 Sep 30;94(20):10594–9.

11. Doyle LA, Yang W, Abruzzo LV, Krogmann T, Gao Y, Rishi AK, u. a. A multidrug resistance transporter from human MCF-7 breast cancer cells. Proc. Natl. Acad. Sci. U.S.A. 1998 Dez 22;95(26):15665–70.

12. Zhou S, Schuetz JD, Bunting KD, Colapietro AM, Sampath J, Morris JJ, u. a. The ABC transporter Bcrp1/ABCG2 is expressed in a wide variety of stem cells and is a molecular determinant of the side-population phenotype. Nat. Med. 2001 Sep;7(9):1028–34.

13. Hirschmann-Jax C, Foster AE, Wulf GG, Nuchtern JG, Jax TW, Gobel U, u. a. A distinct „side population“ of cells with high drug efflux capacity in human tumor cells. Proc. Natl. Acad. Sci. U.S.A. 2004 Sep 28;101(39):14228–33.

14. Aller SG, Yu J, Ward A, Weng Y, Chittaboina S, Zhuo R, u. a. Structure of P-glycoprotein reveals a molecular basis for poly-specific drug binding. Science. 2009 März 27;323(5922):1718–22.

15. Shen S, Callaghan D, Juzwik C, Xiong H, Huang P, Zhang W. ABCG2 reduces ROS-mediated toxicity and inflammation: a potential role in Alzheimer's disease. J. Neurochem. 2010 Sep;114(6):1590–604.

16. Rosenberg MF, Bikadi Z, Chan J, Liu X, Ni Z, Cai X, u. a. The human breast cancer resistance protein (BCRP/ABCG2) shows conformational changes with mitoxantrone. Structure. 2010 März 14;18(4):482–93.

17. Jedlitschky G, Vogelgesang S, Kroemer HK. MDR1-P-glycoprotein (ABCB1)-mediated disposition of amyloid-β peptides: implications for the pathogenesis and therapy of Alzheimer's disease. Clin. Pharmacol. Ther. 2010 Okt;88(4):441–3.

18. Candela P, Gosselet F, Saint-Pol J, Sevin E, Boucau M-C, Boulanger E, u. a. Apical-to-basolateral transport of amyloid-β peptides through blood-brain barrier cells is mediated by the receptor for advanced glycation end-products and is restricted by P-glycoprotein. J. Alzheimers Dis. 2010;22(3):849–59.

19. Rosenberg MF, O'Ryan LP, Hughes G, Zhao Z, Aleksandrov LA, Riordan JR, u. a. The cystic fibrosis transmembrane conductance regulator (CFTR): three-dimensional structure and localization of a channel gate. J. Biol. Chem. 2011 Dez 9;286(49):42647–54.

20. Fletcher JI, Haber M, Henderson MJ, Norris MD. ABC transporters in cancer: more than just drug efflux pumps. Nat. Rev. Cancer. 2010 Feb;10(2):147–56.

21. Colabufo NA, Berardi F, Cantore M, Contino M, Inglese C, Niso M, u. a. Perspectives of P-glycoprotein modulating agents in oncology and neurodegenerative diseases: pharmaceutical, biological, and diagnostic potentials. J. Med. Chem. 2010 März 11;53(5):1883–97.

22. Dean M, Fojo T, Bates S. Tumour stem cells and drug resistance. Nat Rev Cancer. 2005 Apr;5(4):275–84.

23. Goldsborough AS, Handley MD, Dulcey AE, Pluchino KM, Kannan P, Brimacombe KR, u. a. Collateral sensitivity of multidrug-resistant cells to the orphan drug tiopronin. J. Med. Chem. 2011 Juli 28;54(14):4987–97.

24. Giacomini KM, Huang S-M, Tweedie DJ, Benet LZ, Brouwer KLR, Chu X, u. a. Membrane transporters in drug development. Nat Rev Drug Discov. 2010 März;9(3):215–36.

25. Sarkadi B, Homolya L, Szakács G, Váradi A. Human multidrug resistance ABCB and ABCG transporters: participation in a chemoimmunity defense system. Physiol. Rev. 2006 Okt;86(4):1179–236.

26. Bodó A, Bakos E, Szeri F, Váradi A, Sarkadi B. The role of multidrug transporters in drug availability, metabolism and toxicity. Toxicol. Lett. 2003 Apr 11;140-141:133–43.

27. Glavinas H, Krajcsi P, Cserepes J, Sarkadi B. The role of ABC transporters in drug resistance, metabolism and toxicity. Curr Drug Deliv. 2004 Jan;1(1):27–42.

28. Vlaming MLH, Lagas JS, Schinkel AH. Physiological and pharmacological roles of ABCG2 (BCRP): recent findings in Abcg2 knockout mice. Adv. Drug Deliv. Rev. 2009 Jan 31;61(1):14–25.

29. van Herwaarden AE, Schinkel AH. The function of breast cancer resistance protein in epithelial barriers, stem cells and milk secretion of drugs and xenotoxins. Trends Pharmacol. Sci. 2006 Jan;27(1):10–6.

30. Berge KE, Tian H, Graf GA, Yu L, Grishin NV, Schultz J, u. a. Accumulation of dietary cholesterol in sitosterolemia caused by mutations in adjacent ABC transporters. Science. 2000 Dez 1;290(5497):1771–5.

31. MacFarland A, Abramovich DR, Ewen SW, Pearson CK. Stage-specific distribution of P-glycoprotein in first-trimester and full-term human placenta. Histochem. J. 1994 Mai;26(5):417–23.

32. Schinkel AH, Mayer U, Wagenaar E, Mol CA, van Deemter L, Smit JJ, u. a. Normal viability and altered pharmacokinetics in mice lacking mdr1-type (drug-transporting) P-glycoproteins. Proc. Natl. Acad. Sci. U.S.A. 1997 Apr 15;94(8):4028–33.

33. Klimecki WT, Futscher BW, Grogan TM, Dalton WS. P-glycoprotein expression and function in circulating blood cells from normal volunteers. Blood. 1994 Mai 1;83(9):2451–8.

34. Randolph GJ, Beaulieu S, Pope M, Sugawara I, Hoffman L, Steinman RM, u. a. A physiologic function for p-glycoprotein (MDR-1) during the migration of dendritic cells from skin via afferent lymphatic vessels. Proc. Natl. Acad. Sci. U.S.A. 1998 Juni 9;95(12):6924–9.

35. Gupta S, Kim CH, Tsuruo T, Gollapudi S. Preferential expression and activity of multidrug resistance gene 1 product (P-glycoprotein), a functionally active efflux pump, in human CD8+ T cells: a role in cytotoxic effector function. J. Clin. Immunol. 1992 Nov;12(6):451–8.

36. Chong AS, Markham PN, Gebel HM, Bines SD, Coon JS. Diverse multidrug-resistance-modification agents inhibit cytolytic activity of natural killer cells. Cancer Immunol. Immunother. 1993;36(2):133–9.

37. Raghu G, Park SW, Roninson IB, Mechetner EB. Monoclonal antibodies against P-glycoprotein, an MDR1 gene product, inhibit interleukin-2 release from PHA-activated lymphocytes. Exp. Hematol. 1996 Aug;24(10):1258–64.

38. Drach J, Gsur A, Hamilton G, Zhao S, Angerler J, Fiegl M, u. a. Involvement of P-glycoprotein in the transmembrane transport of interleukin-2 (IL-2), IL-4, and interferon-gamma in normal human T lymphocytes. Blood. 1996 Sep 1;88(5):1747–54.

39. Los M, Herr I, Friesen C, Fulda S, Schulze-Osthoff K, Debatin KM. Cross-resistance of CD95- and drug-induced apoptosis as a consequence of deficient activation of caspases (ICE/Ced-3 proteases). Blood. 1997 Okt 15;90(8):3118–29.

40. Bezombes C, Maestre N, Laurent G, Levade T, Bettaïeb A, Jaffrézou JP. Restoration of TNF-alpha-induced ceramide generation and apoptosis in resistant human leukemia KG1a cells by the P-glycoprotein blocker PSC833. FASEB J. 1998 Jan;12(1):101–9.

41. Robinson LJ, Roberts WK, Ling TT, Lamming D, Sternberg SS, Roepe PD. Human MDR 1 protein overexpression delays the apoptotic cascade in Chinese hamster ovary fibroblasts. Biochemistry. 1997 Sep 16;36(37):11169–78.

42. Smyth MJ, Krasovskis E, Sutton VR, Johnstone RW. The drug efflux protein, P-glycoprotein, additionally protects drug-resistant tumor cells from multiple forms of caspase-dependent apoptosis. Proc. Natl. Acad. Sci. U.S.A. 1998 Juni 9;95(12):7024–9.

43. Blagosklonny MV. Treatment with inhibitors of caspases, that are substrates of drug transporters, selectively permits chemotherapy-induced apoptosis in multidrug-resistant cells but protects normal cells. Leukemia. 2001 Juni;15(6):936–41.

44. Johnstone RW, Ruefli AA, Smyth MJ. Multiple physiological functions for multidrug transporter P-glycoprotein? Trends Biochem. Sci. 2000 Jan;25(1):1–6.

45. Johnstone RW, Cretney E, Smyth MJ. P-glycoprotein protects leukemia cells against caspase-dependent, but not caspase-independent, cell death. Blood. 1999 Feb 1;93(3):1075–85.

46. Smit JJ, Schinkel AH, Oude Elferink RP, Groen AK, Wagenaar E, van Deemter L, u. a. Homozygous disruption of the murine mdr2 P-glycoprotein gene leads to a complete absence of phospholipid from bile and to liver disease. Cell. 1993 Nov 5;75(3):451–62.

47. Gerloff T, Stieger B, Hagenbuch B, Madon J, Landmann L, Roth J, u. a. The sister of P-glycoprotein represents the canalicular bile salt export pump of mammalian liver. J. Biol. Chem. 1998 Apr 17;273(16):10046–50.

48. Bodzioch M, Orsó E, Klucken J, Langmann T, Böttcher A, Diederich W, u. a. The gene encoding ATP-binding cassette transporter 1 is mutated in Tangier disease. Nat. Genet. 1999 Aug;22(4):347–51.

49. Brooks-Wilson A, Marcil M, Clee SM, Zhang LH, Roomp K, van Dam M, u. a. Mutations in ABC1 in Tangier disease and familial high-density lipoprotein deficiency. Nat. Genet. 1999 Aug;22(4):336–45.

50. Rust S, Rosier M, Funke H, Real J, Amoura Z, Piette JC, u. a. Tangier disease is caused by mutations in the gene encoding ATP-binding cassette transporter 1. Nat. Genet. 1999 Aug;22(4):352–5.

51. Mack JT, Beljanski V, Soulika AM, Townsend DM, Brown CB, Davis W, u. a. „Skittish“ Abca2 knockout mice display tremor, hyperactivity, and abnormal myelin ultrastructure in the central nervous system. Mol. Cell. Biol. 2007 Jan;27(1):44–53.

52. Sakai H, Tanaka Y, Tanaka M, Ban N, Yamada K, Matsumura Y, u. a. ABCA2 deficiency results in abnormal sphingolipid metabolism in mouse brain. J. Biol. Chem. 2007 Juli 6;282(27):19692–9.

53. Singaraja RR, Visscher H, James ER, Chroni A, Coutinho JM, Brunham LR, u. a. Specific mutations in ABCA1 have discrete effects on ABCA1 function and lipid phenotypes both in vivo and in vitro. Circ. Res. 2006 Aug 18;99(4):389–97.

54. Fitzgerald ML, Morris AL, Rhee JS, Andersson LP, Mendez AJ, Freeman MW. Naturally occurring mutations in the largest extracellular loops of ABCA1 can disrupt its direct interaction with apolipoprotein A-I. J. Biol. Chem. 2002 Sep 6;277(36):33178–87.

55. Sai Y. Biochemical and molecular pharmacological aspects of transporters as determinants of drug disposition. Drug Metab. Pharmacokinet. 2005 Apr;20(2):91–9.

56. Hediger MA, Romero MF, Peng J-B, Rolfs A, Takanaga H, Bruford EA. The ABCs of solute carriers: physiological, pathological and therapeutic implications of human membrane transport proteinsIntroduction. Pflugers Arch. 2004 Feb;447(5):465–8.

57. Koepsell H, Lips K, Volk C. Polyspecific organic cation transporters: structure, function, physiological roles, and biopharmaceutical implications. Pharm. Res. 2007 Juli;24(7):1227–51.

58. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell. 2000 Jan 7;100(1):57–70.

59. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011 März 4;144(5):646–74.

60. Cavallo F, De Giovanni C, Nanni P, Forni G, Lollini P-L. 2011: the immune hallmarks of cancer. Cancer Immunol. Immunother. 2011 März;60(3):319–26.

61. Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. Nat. Med. 2004 Aug;10(8):789–99.

62. Berger AH, Knudson AG, Pandolfi PP. A continuum model for tumour suppression. Nature. 2011 Aug 11;476(7359):163–9.

63. NOWELL PC, HUNGERFORD DA. Chromosome studies on normal and leukemic human leukocytes. J. Natl. Cancer Inst. 1960 Juli;25:85–109.

64. Knudson AG Jr. Mutation and cancer: statistical study of retinoblastoma. Proc. Natl. Acad. Sci. U.S.A. 1971 Apr;68(4):820–3.

65. Nussbaumer S, Bonnabry P, Veuthey J-L, Fleury-Souverain S. Analysis of anticancer drugs: a review. Talanta. 2011 Okt 15;85(5):2265–89.

66. McDermott U, Settleman J. Personalized cancer therapy with selective kinase inhibitors: an emerging paradigm in medical oncology. J. Clin. Oncol. 2009 Nov 20;27(33):5650–9.

67. Martini M, Vecchione L, Siena S, Tejpar S, Bardelli A. Targeted therapies: how personal should we go? Nat Rev Clin Oncol. 2011;9(2):87–97.

68. Gottesman MM, Ling V. The molecular basis of multidrug resistance in cancer: the early years of P-glycoprotein research. FEBS Lett. 2006 Feb 13;580(4):998–1009.

69. Gottesman MM, Fojo T, Bates SE. Multidrug resistance in cancer: role of ATP-dependent transporters. Nat Rev Cancer. 2002 Jan;2(1):48–58.

70. Dean M. The genetics of ATP-binding cassette transporters. Meth. Enzymol. 2005;400:409–29.

71. Scripture CD, Figg WD. Drug interactions in cancer therapy. Nat Rev Cancer. 2006 Juli;6(7):546–58.

72. Licht T, Goldenberg SK, Vieira WD, Gottesman MM, Pastan I. Drug selection of MDR1-transduced hematopoietic cells ex vivo increases transgene expression and chemoresistance in reconstituted bone marrow in mice. Gene Ther. 2000 Feb;7:348–58.

73. Chen KG, Valencia JC, Gillet J-P, Hearing VJ, Gottesman MM. Involvement of ABC transporters in melanogenesis and the development of multidrug resistance of melanoma. Pigment Cell Melanoma Res. 2009 Dez;22(6):740–9.

74. Gillet J-P, Efferth T, Remacle J. Chemotherapy-induced resistance by ATP-binding cassette transporter genes. Biochim. Biophys. Acta. 2007 Juni;1775(2):237–62.

75. Gillet J-P, Efferth T, Steinbach D, Hamels J, de Longueville F, Bertholet V, u. a. Microarray-based detection of multidrug resistance in human tumor cells by expression profiling of ATP-binding cassette transporter genes. Cancer Res. 2004 Dez 15;64(24):8987–93.

76. Gillet J-P, Wang J, Calcagno AM, Green LJ, Varma S, Bunkholt Elstrand M, u. a. Clinical Relevance of Multidrug Resistance Gene Expression in Ovarian Serous Carcinoma Effusions. Molecular Pharmaceutics [Internet]. 2011 Juli 15 [zitiert 2011 Okt 2]; Available von: http://www.ncbi.nlm.nih.gov/pubmed/21761824

77. Ambudkar SV, Dey S, Hrycyna CA, Ramachandra M, Pastan I, Gottesman MM. Biochemical, cellular, and pharmacological aspects of the multidrug transporter. Annu. Rev. Pharmacol. Toxicol. 1999;39:361–98.

78. Hegedus C, Ozvegy-Laczka C, Szakács G, Sarkadi B. Interaction of ABC multidrug transporters with anticancer protein kinase inhibitors: substrates and/or inhibitors? Curr Cancer Drug Targets. 2009 Mai;9(3):252–72.

79. Hegedus C, Ozvegy-Laczka C, Apáti A, Magócsi M, Német K, Orfi L, u. a. Interaction of nilotinib, dasatinib and bosutinib with ABCB1 and ABCG2: implications for altered anti-cancer effects and pharmacological properties. Br. J. Pharmacol. 2009 Okt;158(4):1153–64.

80. Norris MD, De Graaf D, Haber M, Kavallaris M, Madafiglio J, Gilbert J, u. a. Involvement of MDR1 P-glycoprotein in multifactorial resistance to methotrexate. Int. J. Cancer. 1996 März 1;65(5):613–9.

81. Consoli U, Van NT, Neamati N, Mahadevia R, Beran M, Zhao S, u. a. Cellular pharmacology of mitoxantrone in p-glycoprotein-positive and -negative human myeloid leukemic cell lines. Leukemia. 1997 Dez;11(12):2066–74.

82. Krishna R, Mayer LD. Multidrug resistance (MDR) in cancer. Mechanisms, reversal using modulators of MDR and the role of MDR modulators in influencing the pharmacokinetics of anticancer drugs. Eur J Pharm Sci. 2000 Okt;11(4):265–83.

83. Thomas H, Coley HM. Overcoming multidrug resistance in cancer: an update on the clinical strategy of inhibiting p-glycoprotein. Cancer Control. 2003 Apr;10(2):159–65.

84. Hu XF, Nadalin G, De Luise M, Martin TJ, Wakeling A, Huggins R, u. a. Circumvention of doxorubicin resistance in multi-drug resistant human leukaemia and lung cancer cells by the pure antioestrogen ICI 164384. Eur. J. Cancer. 1991;27(6):773–7.

85. Stupp R, Bauer J, Pagani O, Gerard B, Cerny T, Sessa C, u. a. Ventricular arrhythmia and torsade de pointe: dose limiting toxicities of the MDR-modulator S9788 in a phase I trial. Ann. Oncol. 1998 Nov;9(11):1233–42.

86. Wandel C, Kim RB, Kajiji S, Guengerich P, Wilkinson GR, Wood AJ. P-glycoprotein and cytochrome P-450 3A inhibition: dissociation of inhibitory potencies. Cancer Res. 1999 Aug 15;59(16):3944–8.

87. Minderman H, O'Loughlin KL, Pendyala L, Baer MR. VX-710 (biricodar) increases drug retention and enhances chemosensitivity in resistant cells overexpressing P-glycoprotein, multidrug resistance protein, and breast cancer resistance protein. Clin. Cancer Res. 2004 März 1;10(5):1826–34.

88. Martin C, Higgins CF, Callaghan R. The vinblastine binding site adopts high- and low-affinity conformations during a transport cycle of P-glycoprotein. Biochemistry. 2001 Dez 25;40(51):15733–42.

89. Martin C, Berridge G, Higgins CF, Mistry P, Charlton P, Callaghan R. Communication between multiple drug binding sites on P-glycoprotein. Mol. Pharmacol. 2000 Sep;58(3):624–32.

90. Pleban K, Kopp S, Csaszar E, Peer M, Hrebicek T, Rizzi A, u. a. P-glycoprotein substrate binding domains are located at the transmembrane domain/transmembrane domain interfaces: a combined photoaffinity labeling-protein homology modeling approach. Mol. Pharmacol. 2005 Feb;67(2):365–74.

91. Nobili S, Landini I, Giglioni B, Mini E. Pharmacological strategies for overcoming multidrug resistance. Curr Drug Targets. 2006 Juli;7(7):861–79.

92. Fox E, Bates SE. Tariquidar (XR9576): a P-glycoprotein drug efflux pump inhibitor. Expert Rev Anticancer Ther. 2007 Apr;7(4):447–59.

93. Rubin EH, de Alwis DP, Pouliquen I, Green L, Marder P, Lin Y, u. a. A phase I trial of a potent P-glycoprotein inhibitor, Zosuquidar.3HCl trihydrochloride (LY335979), administered orally in

combination with doxorubicin in patients with advanced malignancies. Clin. Cancer Res. 2002 Dez;8(12):3710–7.

94. Kuppens IELM, Witteveen EO, Jewell RC, Radema SA, Paul EM, Mangum SG, u. a. A phase I, randomized, open-label, parallel-cohort, dose-finding study of elacridar (GF120918) and oral topotecan in cancer patients. Clin. Cancer Res. 2007 Juni 1;13(11):3276–85.

95. Dörner B, Kuntner C, Bankstahl JP, Bankstahl M, Stanek J, Wanek T, u. a. Synthesis and small-animal positron emission tomography evaluation of [11C]-elacridar as a radiotracer to assess the distribution of P-glycoprotein at the blood-brain barrier. J. Med. Chem. 2009 Okt 8;52(19):6073–82.

96. Kuppens IELM, Bosch TM, van Maanen MJ, Rosing H, Fitzpatrick A, Beijnen JH, u. a. Oral bioavailability of docetaxel in combination with OC144-093 (ONT-093). Cancer Chemother. Pharmacol. 2005 Jan;55(1):72–8.

97. Pleban K, Kaiser D, Kopp S, Peer M, Chiba P, Ecker GF. Targeting drug-efflux pumps -- a pharmacoinformatic approach. Acta Biochim. Pol. 2005;52(3):737–40.

98. Klepsch F, Stockner T, Erker T, Muller M, Chiba P, Ecker GF. Using structural and mechanistic information to design novel inhibitors/substrates of P-glycoprotein. Curr Top Med Chem. 2010;10:1769–74.

99. Klepsch F, Jabeen I, Chiba P, Ecker GF. Pharmacoinformatic approaches to design natural product type ligands of ABC-transporters. Curr. Pharm. Des. 2010 Mai;16(15):1742–52.

100. Cox CD, Breslin MJ, Whitman DB, Coleman PJ, Garbaccio RM, Fraley ME, u. a. Kinesin spindle protein (KSP) inhibitors. Part V: discovery of 2-propylamino-2,4-diaryl-2,5-dihydropyrroles as potent, water-soluble KSP inhibitors, and modulation of their basicity by beta-fluorination to overcome cellular efflux by P-glycoprotein. Bioorg. Med. Chem. Lett. 2007 Mai 15;17(10):2697–702.

101. Dunne G, Breen L, Collins DM, Roche S, Clynes M, O'Connor R. Modulation of P-gp expression by lapatinib. Invest New Drugs [Internet]. 2010 Juli 6; Available von: http://www.ncbi.nlm.nih.gov/pubmed/20607587

102. Ozvegy-Laczka C, Cserepes J, Elkind NB, Sarkadi B. Tyrosine kinase inhibitor resistance in cancer: role of ABC multidrug transporters. Drug Resist. Updat. 2005 Apr;8(1-2):15–26.

103. Leggas M, Panetta JC, Zhuang Y, Schuetz JD, Johnston B, Bai F, u. a. Gefitinib modulates the function of multiple ATP-binding cassette transporters in vivo. Cancer Res. 2006 Mai 1;66(9):4802–7.

104. Filipits M, Suchomel RW, Dekan G, Haider K, Valdimarsson G, Depisch D, u. a. MRP and MDR1 gene expression in primary breast carcinomas. Clin. Cancer Res. 1996 Juli;2(7):1231–7.

105. Nooter K, de la Riviere GB, Klijn J, Stoter G, Foekens J. Multidrug resistance protein in recurrent breast cancer. Lancet. 1997 Juni 28;349(9069):1885–6.

106. Wright SR, Boag AH, Valdimarsson G, Hipfner DR, Campling BG, Cole SP, u. a. Immunohistochemical detection of multidrug resistance protein in human lung cancer and normal lung. Clin. Cancer Res. 1998 Sep;4(9):2279–89.

107. Lautier D, Canitrot Y, Deeley RG, Cole SP. Multidrug resistance mediated by the multidrug resistance protein (MRP) gene. Biochem. Pharmacol. 1996 Okt 11;52(7):967–77.

108. Deeley RG, Westlake C, Cole SPC. Transmembrane transport of endo- and xenobiotics by mammalian ATP-binding cassette multidrug resistance proteins. Physiol. Rev. 2006 Juli;86(3):849–99.

109. Morrow CS, Peklak-Scott C, Bishwokarma B, Kute TE, Smitherman PK, Townsend AJ. Multidrug resistance protein 1 (MRP1, ABCC1) mediates resistance to mitoxantrone via glutathione-dependent drug efflux. Mol. Pharmacol. 2006 Apr;69(4):1499–505.

110. Chu XY, Suzuki H, Ueda K, Kato Y, Akiyama S, Sugiyama Y. Active efflux of CPT-11 and its metabolites in human KB-derived cell lines. J. Pharmacol. Exp. Ther. 1999 Feb;288(2):735–41.

111. Bakos E, Evers R, Sinkó E, Váradi A, Borst P, Sarkadi B. Interactions of the human multidrug resistance proteins MRP1 and MRP2 with organic anions. Mol. Pharmacol. 2000 Apr;57(4):760–8.

112. Allikmets R, Schriml LM, Hutchinson A, Romano-Spica V, Dean M. A human placenta-specific ATP-binding cassette gene (ABCP) on chromosome 4q22 that is involved in multidrug resistance. Cancer Res. 1998 Dez 1;58(23):5337–9.

113. Polgar O, Ozvegy-Laczka C, Robey RW, Morisaki K, Okada M, Tamaki A, u. a. Mutational studies of G553 in TM5 of ABCG2: a residue potentially involved in dimerization. Biochemistry. 2006 Apr 25;45(16):5251–60.

114. Yoh K, Ishii G, Yokose T, Minegishi Y, Tsuta K, Goto K, u. a. Breast cancer resistance protein impacts clinical outcome in platinum-based chemotherapy for advanced non-small cell lung cancer. Clin. Cancer Res. 2004 März;10:1691–7.

115. Steinbach D, Sell W, Voigt A, Hermann J, Zintl F, Sauerbrey A. BCRP gene expression is associated with a poor response to remission induction therapy in childhood acute myeloid leukemia. Leukemia. 2002 Aug;16(8):1443–7.

116. van den Heuvel-Eibrink MM, Wiemer EAC, Prins A, Meijerink JPP, Vossebeld PJM, van der Holt B, u. a. Increased expression of the breast cancer resistance protein (BCRP) in relapsed or refractory acute myeloid leukemia (AML). Leukemia. 2002 Mai;16(5):833–9.

117. Wang X-K, Fu L-W. Interaction of tyrosine kinase inhibitors with the MDR- related ABC transporter proteins. Curr. Drug Metab. 2010 Sep;11(7):618–28.

118. Vulevic B, Chen Z, Boyd JT, Davis W Jr, Walsh ES, Belinsky MG, u. a. Cloning and characterization of human adenosine 5'-triphosphate-binding cassette, sub-family A, transporter 2 (ABCA2). Cancer Res. 2001 Apr 15;61(8):3339–47.

119. Chapuy B, Koch R, Radunski U, Corsham S, Cheong N, Inagaki N, u. a. Intracellular ABC transporter A3 confers multidrug resistance in leukemia cells by lysosomal drug sequestration. Leukemia. 2008 Aug;22(8):1576–86.

120. Smith AJ, van Helvoort A, van Meer G, Szabo K, Welker E, Szakacs G, u. a. MDR3 P-glycoprotein, a phosphatidylcholine translocase, transports several cytotoxic drugs and directly

interacts with drugs as judged by interference with nucleotide trapping. J. Biol. Chem. 2000 Aug 4;275(31):23530–9.

121. Frank NY, Margaryan A, Huang Y, Schatton T, Waaga-Gasser AM, Gasser M, u. a. ABCB5-mediated doxorubicin transport and chemoresistance in human malignant melanoma. Cancer Res. 2005 Mai 15;65(10):4320–33.

122. Childs S, Yeh RL, Hui D, Ling V. Taxol resistance mediated by transfection of the liver-specific sister gene of P-glycoprotein. Cancer Res. 1998 Sep 15;58(18):4160–7.

123. Zelcer N, Saeki T, Reid G, Beijnen JH, Borst P. Characterization of drug transport by the human multidrug resistance protein 3 (ABCC3). J. Biol. Chem. 2001 Dez 7;276(49):46400–7.

124. Ritter CA, Jedlitschky G, Meyer zu Schwabedissen H, Grube M, Köck K, Kroemer HK. Cellular export of drugs and signaling molecules by the ATP-binding cassette transporters MRP4 (ABCC4) and MRP5 (ABCC5). Drug Metab. Rev. 2005;37(1):253–78.

125. Belinsky MG, Chen Z-S, Shchaveleva I, Zeng H, Kruh GD. Characterization of the drug resistance and transport properties of multidrug resistance protein 6 (MRP6, ABCC6). Cancer Res. 2002 Nov 1;62(21):6172–7.

126. Hopper-Borge E, Chen Z-S, Shchaveleva I, Belinsky MG, Kruh GD. Analysis of the drug resistance profile of multidrug resistance protein 7 (ABCC10): resistance to docetaxel. Cancer Res. 2004 Juli 15;64(14):4927–30.

127. Chen Z-S, Guo Y, Belinsky MG, Kotova E, Kruh GD. Transport of bile acids, sulfated steroids, estradiol 17-beta-D-glucuronide, and leukotriene C4 by human multidrug resistance protein 8 (ABCC11). Mol. Pharmacol. 2005 Feb;67(2):545–57.

128. Borst P, Evers R, Kool M, Wijnholds J. A family of drug transporters: the multidrug resistance-associated proteins. J. Natl. Cancer Inst. 2000 Aug 16;92(16):1295–302.

129. Visvader JE, Lindeman GJ. Cancer stem cells in solid tumours: accumulating evidence and unresolved questions. Nat Rev Cancer. 2008 Okt;8(10):755–68.

130. Ailles LE, Weissman IL. Cancer stem cells in solid tumors. Curr. Opin. Biotechnol. 2007 Okt;18(5):460–6.

131. Reya T, Morrison SJ, Clarke MF, Weissman IL. Stem cells, cancer, and cancer stem cells. Nature. 2001 Nov 1;414(6859):105–11.

132. Wang JCY. Evaluating therapeutic efficacy against cancer stem cells: new challenges posed by a new paradigm. Cell Stem Cell. 2007 Nov;1(5):497–501.

133. Sarkar FH, Li Y, Wang Z, Kong D. Pancreatic cancer stem cells and EMT in drug resistance and metastasis. Minerva Chir. 2009 Okt;64(5):489–500.

134. Hirschmann-Jax C, Foster AE, Wulf GG, Goodell MA, Brenner MK. A distinct „side population“ of cells in human tumor cells: implications for tumor biology and therapy. Cell Cycle. 2005 Feb;4(2):203–5.

135. Schatton T, Murphy GF, Frank NY, Yamaura K, Waaga-Gasser AM, Gasser M, u. a. Identification of cells initiating human melanomas. Nature. 2008 Jan 17;451(7176):345–9.

136. Angelastro JM, Lamé MW. Overexpression of CD133 promotes drug resistance in C6 glioma cells. Mol. Cancer Res. 2010 Aug;8(8):1105–15.

137. Shervington A, Lu C. Expression of multidrug resistance genes in normal and cancer stem cells. Cancer Invest. 2008 Juni;26(5):535–42.

138. Calcagno AM, Salcido CD, Gillet J-P, Wu C-P, Fostel JM, Mumau MD, u. a. Prolonged drug selection of breast cancer cells and enrichment of cancer stem cell characteristics. J. Natl. Cancer Inst. 2010 Nov 3;102(21):1637–52.

139. Bertolini G, Roz L, Perego P, Tortoreto M, Fontanella E, Gatti L, u. a. Highly tumorigenic lung cancer CD133+ cells display stem-like features and are spared by cisplatin treatment. Proc. Natl. Acad. Sci. U.S.A. 2009 Sep 22;106(38):16281–6.

140. Creighton CJ, Li X, Landis M, Dixon JM, Neumeister VM, Sjolund A, u. a. Residual breast cancers after conventional therapy display mesenchymal as well as tumor-initiating features. Proc. Natl. Acad. Sci. U.S.A. 2009 Aug 18;106(33):13820–5.

141. Fillmore CM, Kuperwasser C. Human breast cancer cell lines contain stem-like cells that self-renew, give rise to phenotypically diverse progeny and survive chemotherapy. Breast Cancer Res. 2008;10(2):R25.

142. Singh A, Settleman J. EMT, cancer stem cells and drug resistance: an emerging axis of evil in the war on cancer. Oncogene. 2010 Aug 26;29(34):4741–51.

143. Xia X, Yang J, Li F, Li Y, Zhou X, Dai Y, u. a. Image-based chemical screening identifies drug efflux inhibitors in lung cancer cells. Cancer Res. 2010 Okt 1;70(19):7723–33.

144. Visvader JE. Cells of origin in cancer. Nature. 2011 Jan 20;469(7330):314–22.

145. Vogelgesang S, Cascorbi I, Schroeder E, Pahnke J, Kroemer HK, Siegmund W, u. a. Deposition of Alzheimer's beta-amyloid is inversely correlated with P-glycoprotein expression in the brains of elderly non-demented humans. Pharmacogenetics. 2002 Okt;12(7):535–41.

146. Vogelgesang S, Warzok RW, Cascorbi I, Kunert-Keil C, Schroeder E, Kroemer HK, u. a. The role of P-glycoprotein in cerebral amyloid angiopathy; implications for the early pathogenesis of Alzheimer's disease. Curr Alzheimer Res. 2004 Mai;1(2):121–5.

147. Droździk M, Białecka M, Myśliwiec K, Honczarenko K, Stankiewicz J, Sych Z. Polymorphism in the P-glycoprotein drug transporter MDR1 gene: a possible link between environmental and genetic factors in Parkinson's disease. Pharmacogenetics. 2003 Mai;13(5):259–63.

148. Kortekaas R, Leenders KL, van Oostrom JCH, Vaalburg W, Bart J, Willemsen ATM, u. a. Blood-brain barrier dysfunction in parkinsonian midbrain in vivo. Ann. Neurol. 2005 Feb;57(2):176–9.

149. Bartels AL, Kortekaas R, Bart J, Willemsen ATM, de Klerk OL, de Vries JJ, u. a. Blood-brain barrier P-glycoprotein function decreases in specific brain regions with aging: a possible role in progressive neurodegeneration. Neurobiol. Aging. 2009 Nov;30(11):1818–24.

150. Martel F, Calhau C, Soares-da-Silva P, Azevedo I. Transport of [3H]MPP+ in an immortalized rat brain microvessel endothelial cell line (RBE 4). Naunyn Schmiedebergs Arch. Pharmacol. 2001 Jan;363(1):1–10.

151. Kuhnke D, Jedlitschky G, Grube M, Krohn M, Jucker M, Mosyagin I, u. a. MDR1-P-Glycoprotein (ABCB1) Mediates Transport of Alzheimer's amyloid-beta peptides--implications for the mechanisms of Abeta clearance at the blood-brain barrier. Brain Pathol. 2007 Okt;17(4):347–53.

152. Robey RW, Lazarowski A, Bates SE. P-glycoprotein--a clinical target in drug-refractory epilepsy? Mol. Pharmacol. 2008 Mai;73(5):1343–6.

153. Bates SF, Chen C, Robey R, Kang M, Figg WD, Fojo T. Reversal of multidrug resistance: lessons from clinical oncology. Novartis Found. Symp. 2002;243:83–96; discussion 96–102, 180–185.

154. Löscher W. How to explain multidrug resistance in epilepsy? Epilepsy Curr. 2005 Juni;5(3):107–12.

155. Siddiqui A, Kerb R, Weale ME, Brinkmann U, Smith A, Goldstein DB, u. a. Association of multidrug resistance in epilepsy with a polymorphism in the drug-transporter gene ABCB1. N. Engl. J. Med. 2003 Apr 10;348(15):1442–8.

156. Sisodiya SM, Bates SE. Treatment of drug resistance in epilepsy: one step at a time. Lancet Neurol. 2006 Mai;5(5):380–1.

157. Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, Grzelczak Z, u. a. Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. Science. 1989 Sep 8;245(4922):1066–73.

158. Antigny F, Norez C, Becq F, Vandebrouck C. CFTR and Ca Signaling in Cystic Fibrosis. Front Pharmacol. 2011;2:67.

159. Aguilar-Bryan L, Nichols CG, Wechsler SW, Clement JP 4th, Boyd AE 3rd, González G, u. a. Cloning of the beta cell high-affinity sulfonylurea receptor: a regulator of insulin secretion. Science. 1995 Apr 21;268(5209):423–6.

160. Seino S, Shibasaki T, Minami K. Dynamics of insulin secretion and the clinical implications for obesity and diabetes. J. Clin. Invest. 2011 Juni;121(6):2118–25.

161. Pohl A, Devaux PF, Herrmann A. Function of prokaryotic and eukaryotic ABC proteins in lipid transport. Biochim. Biophys. Acta. 2005 März 21;1733(1):29–52.

162. Cohen ML. Epidemiology of drug resistance: implications for a post-antimicrobial era. Science. 1992 Aug 21;257(5073):1050–5.

163. Culliton BJ. Drug-resistant TB may bring epidemic. Nature. 1992 Apr 9;356(6369):473.

164. Lambert G, Estévez-Salmeron L, Oh S, Liao D, Emerson BM, Tlsty TD, u. a. An analogy between the evolution of drug resistance in bacterial communities and malignant tissues. Nat. Rev. Cancer. 2011 Mai;11(5):375–82.

165. Pagès J-M, Amaral L, Fanning S. An original deal for new molecule: reversal of efflux pump activity, a rational strategy to combat gram-negative resistant bacteria. Curr. Med. Chem. 2011;18(19):2969–80.

166. Rameis H. Quinidine-digoxin interaction: are the pharmacokinetics of both drugs altered? Int J Clin Pharmacol Ther Toxicol. 1985 März;23(3):145–53.

167. Fromm MF, Kim RB, Stein CM, Wilkinson GR, Roden DM. Inhibition of P-glycoprotein-mediated drug transport: A unifying mechanism to explain the interaction between digoxin and quinidine [seecomments]. Circulation. 1999 Feb 2;99(4):552–7.

168. Polli JW, Wring SA, Humphreys JE, Huang L, Morgan JB, Webster LO, u. a. Rational use of in vitro P-glycoprotein assays in drug discovery. J. Pharmacol. Exp. Ther. 2001 Nov;299:620–8.

169. Ding R, Tayrouz Y, Riedel K-D, Burhenne J, Weiss J, Mikus G, u. a. Substantial pharmacokinetic interaction between digoxin and ritonavir in healthy volunteers. Clin. Pharmacol. Ther. 2004 Juli;76(1):73–84.

170. Drescher S, Glaeser H, Mürdter T, Hitzl M, Eichelbaum M, Fromm MF. P-glycoprotein-mediated intestinal and biliary digoxin transport in humans. Clin. Pharmacol. Ther. 2003 März;73(3):223–31.

171. Eberl S, Renner B, Neubert A, Reisig M, Bachmakov I, König J, u. a. Role of p-glycoprotein inhibition for drug interactions: evidence from in vitro and pharmacoepidemiological studies. Clin Pharmacokinet. 2007;46(12):1039–49.

172. Rengelshausen J, Göggelmann C, Burhenne J, Riedel K-D, Ludwig J, Weiss J, u. a. Contribution of increased oral bioavailability and reduced nonglomerular renal clearance of digoxin to the digoxin-clarithromycin interaction. Br J Clin Pharmacol. 2003 Juli;56(1):32–8.

173. Sakaeda T, Nakamura T, Horinouchi M, Kakumoto M, Ohmoto N, Sakai T, u. a. MDR1 genotype-related pharmacokinetics of digoxin after single oral administration in healthy Japanese subjects. Pharm. Res. 2001 Okt;18(10):1400–4.

174. Verstuyft C, Schwab M, Schaeffeler E, Kerb R, Brinkmann U, Jaillon P, u. a. Digoxin pharmacokinetics and MDR1 genetic polymorphisms. Eur. J. Clin. Pharmacol. 2003 Apr;58(12):809–12.

175. Anglicheau D, Thervet E, Etienne I, Hurault De Ligny B, Le Meur Y, Touchard G, u. a. CYP3A5 and MDR1 genetic polymorphisms and cyclosporine pharmacokinetics after renal transplantation. Clin. Pharmacol. Ther. 2004 Mai;75(5):422–33.

176. Arboix M, Paz OG, Colombo T, D'Incalci M. Multidrug resistance-reversing agents increase vinblastine distribution in normal tissues expressing the P-glycoprotein but do not enhance drug penetration in brain and testis. J. Pharmacol. Exp. Ther. 1997 Juni;281(3):1226–30.

177. Spahn-Langguth H, Baktir G, Radschuweit A, Okyar A, Terhaag B, Ader P, u. a. P-glycoprotein transporters and the gastrointestinal tract: evaluation of the potential in vivo relevance of in vitro data employing talinolol as model compound. Int J Clin Pharmacol Ther. 1998 Jan;36(1):16–24.

178. Schwarz UI, Gramatté T, Krappweis J, Oertel R, Kirch W. P-glycoprotein inhibitor erythromycin increases oral bioavailability of talinolol in humans. Int J Clin Pharmacol Ther. 2000 Apr;38(4):161–7.

179. Pauli-Magnus C, Rekersbrink S, Klotz U, Fromm MF. Interaction of omeprazole, lansoprazole and pantoprazole with P-glycoprotein. Naunyn Schmiedebergs Arch. Pharmacol. 2001 Dez;364:551–7.

180. Blume H, Donath F, Warnke A, Schug BS. Pharmacokinetic drug interaction profiles of proton pump inhibitors. Drug Saf. 2006;29:769–84.

181. Sipe BE, Jones RJ, Bokhart GH. Rhabdomyolysis causing AV blockade due to possible atorvastatin, esomeprazole, and clarithromycin interaction. Ann Pharmacother. 2003 Juni;37:808–11.

182. Vaz RJ, Klabunde. Antitargets-Prediction and Prevention of Drug Side Effects. Weinheim: Wiley-VCH; 2008.

183. van Waterschoot RA, Schinkel AH. A critical analysis of the interplay between cytochrome P450 3A and P-glycoprotein: recent insights from knockout and transgenic mice. Pharmacol. Rev. 2011 Juni;63:390–410.

184. Ekroos M, Sjögren T. Structural basis for ligand promiscuity in cytochrome P450 3A4. Proc. Natl. Acad. Sci. U.S.A. 2006 Sep 12;103(37):13682–7.

185. Urquhart BL, Tirona RG, Kim RB. Nuclear receptors and the regulation of drug-metabolizing enzymes and drug transporters: implications for interindividual variability in response to drugs. J Clin Pharmacol. 2007 Mai;47(5):566–78.

186. Kivistö KT, Niemi M, Fromm MF. Functional interaction of intestinal CYP3A4 and P-glycoprotein. Fundam Clin Pharmacol. 2004 Dez;18(6):621–6.

187. Veronese ML, Sun W, Giantonio B, Berlin J, Shults J, Davis L, u. a. A phase II trial of gefitinib with 5-fluorouracil, leucovorin, and irinotecan in patients with colorectal cancer. Br. J. Cancer. 2005 Mai 23;92(10):1846–9.

188. Ozvegy-Laczka C, Hegedus T, Várady G, Ujhelly O, Schuetz JD, Váradi A, u. a. High-affinity interaction of tyrosine kinase inhibitors with the ABCG2 multidrug transporter. Mol. Pharmacol. 2004 Juni;65(6):1485–95.

189. Wierdl M, Wall A, Morton CL, Sampath J, Danks MK, Schuetz JD, u. a. Carboxylesterase-mediated sensitization of human tumor cells to CPT-11 cannot override ABCG2-mediated drug resistance. Mol. Pharmacol. 2003 Aug;64(2):279–88.

190. Rees DC, Johnson E, Lewinson O. ABC transporters: the power to change. Nat. Rev. Mol. Cell Biol. 2009 März;10:218–27.

191. Loo TW, Bartlett MC, Clarke DM. Identification of residues in the drug translocation pathway of the human multidrug resistance P-glycoprotein by arginine mutagenesis. J. Biol. Chem. 2009 Sep 4;284(36):24074–87.

192. Loo TW, Bartlett MC, Clarke DM. Nucleotide binding, ATP hydrolysis, and mutation of the catalytic carboxylates of human P-glycoprotein cause distinct conformational changes in the transmembrane segments. Biochemistry. 2007 Aug 14;46(32):9328–36.

193. Loo TW, Bartlett MC, Clarke DM. Suppressor mutations in the transmembrane segments of P-glycoprotein promote maturation of processing mutants and disrupt a subset of drug-binding sites. J. Biol. Chem. 2007 Nov 2;282(44):32043–52.

194. Loo TW, Bartlett MC, Clarke DM. Disulfide cross-linking analysis shows that transmembrane segments 5 and 8 of human P-glycoprotein are close together on the cytoplasmic side of the membrane. J. Biol. Chem. 2004 Feb 27;279(9):7692–7.

195. Loo TW, Bartlett MC, Clarke DM. Methanethiosulfonate derivatives of rhodamine and verapamil activate human P-glycoprotein at different sites. J. Biol. Chem. 2003 Dez 12;278(50):50136–41.

196. Loo TW, Bartlett MC, Clarke DM. Rescue of folding defects in ABC transporters using pharmacological chaperones. J. Bioenerg. Biomembr. 2005 Dez;37(6):501–7.

197. Ward A, Mulligan S, Carragher B, Chang G, Milligan RA. Nucleotide dependent packing differences in helical crystals of the ABC transporter MsbA. J. Struct. Biol. 2009 März;165(3):169–75.

198. Tombline G, Bartholomew LA, Tyndall GA, Gimi K, Urbatsch IL, Senior AE. Properties of P-glycoprotein with mutations in the „catalytic carboxylate“ glutamate residues. J. Biol. Chem. 2004 Nov 5;279(45):46518–26.

199. Rosenberg MF, Kamis AB, Callaghan R, Higgins CF, Ford RC. Three-dimensional structures of the mammalian multidrug resistance P-glycoprotein demonstrate major conformational changes in the transmembrane domains upon nucleotide binding. J. Biol. Chem. 2003 März 7;278(10):8294–9.

200. Rosenberg MF, Callaghan R, Ford RC, Higgins CF. Structure of the multidrug resistance P-glycoprotein to 2.5 nm resolution determined by electron microscopy and image analysis. J. Biol. Chem. 1997 Apr 18;272(16):10685–94.

201. Locher KP. Review. Structure and mechanism of ATP-binding cassette transporters. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 2009 Jan 27;364(1514):239–45.

202. Dawson RJP, Locher KP. Structure of the multidrug ABC transporter Sav1866 from Staphylococcus aureus in complex with AMP-PNP. FEBS Lett. 2007 März 6;581(5):935–8.

203. Rothnie A, Storm J, McMahon R, Taylor A, Kerr ID, Callaghan R. The coupling mechanism of P-glycoprotein involves residue L339 in the sixth membrane spanning segment. FEBS Lett. 2005 Juli 18;579(18):3984–90.

204. Rothnie A, Storm J, Campbell J, Linton KJ, Kerr ID, Callaghan R. The topography of transmembrane segment six is altered during the catalytic cycle of P-glycoprotein. J. Biol. Chem. 2004 Aug 13;279(33):34913–21.

205. Loo TW, Bartlett MC, Clarke DM. ATP hydrolysis promotes interactions between the extracellular ends of transmembrane segments 1 and 11 of human multidrug resistance P-glycoprotein. Biochemistry. 2005 Aug 2;44(30):10250–8.

206. Ecker GF, Pleban K, Kopp S, Csaszar E, Poelarends GJ, Putman M, u. a. A three-dimensional model for the substrate binding domain of the multidrug ATP binding cassette transporter LmrA. Mol. Pharmacol. 2004 Nov;66(5):1169–79.

207. Sonveaux N, Vigano C, Shapiro AB, Ling V, Ruysschaert JM. Ligand-mediated tertiary structure changes of reconstituted P-glycoprotein. A tryptophan fluorescence quenching analysis. J. Biol. Chem. 1999 Juni 18;274(25):17649–54.

208. Neumann L, Abele R, Tampé R. Thermodynamics of peptide binding to the transporter associated with antigen processing (TAP). J. Mol. Biol. 2002 Dez 13;324(5):965–73.

209. Petronilli V, Ames GF. Binding protein-independent histidine permease mutants. Uncoupling of ATP hydrolysis from transmembrane signaling. J. Biol. Chem. 1991 Sep 5;266(25):16293–6.

210. Liu R, Sharom FJ. Site-directed fluorescence labeling of P-glycoprotein on cysteine residues in the nucleotide binding domains. Biochemistry. 1996 Sep 10;35(36):11865–73.

211. Gabriel MP, Storm J, Rothnie A, Taylor AM, Linton KJ, Kerr ID, u. a. Communication between the nucleotide binding domains of P-glycoprotein occurs via conformational changes that involve residue 508. Biochemistry. 2003 Juli 1;42(25):7780–9.

212. Rosenberg MF, Velarde G, Ford RC, Martin C, Berridge G, Kerr ID, u. a. Repacking of the transmembrane domains of P-glycoprotein during the transport ATPase cycle. EMBO J. 2001 Okt 15;20(20):5615–25.

213. van Veen HW, Margolles A, Müller M, Higgins CF, Konings WN. The homodimeric ATP-binding cassette transporter LmrA mediates multidrug transport by an alternating two-site (two-cylinder engine) mechanism. EMBO J. 2000 Juni 1;19(11):2503–14.

214. Urbatsch IL, Tyndall GA, Tombline G, Senior AE. P-glycoprotein catalytic mechanism: studies of the ADP-vanadate inhibited state. J. Biol. Chem. 2003 Juni 20;278(25):23171–9.

215. Urbatsch IL, Sankaran B, Weber J, Senior AE. P-glycoprotein is stably inhibited by vanadate-induced trapping of nucleotide at a single catalytic site. J. Biol. Chem. 1995 Aug 18;270(33):19383–90.

216. Verdon G, Albers SV, Dijkstra BW, Driessen AJM, Thunnissen AMWH. Crystal structures of the ATPase subunit of the glucose ABC transporter from Sulfolobus solfataricus: nucleotide-free and nucleotide-bound conformations. J. Mol. Biol. 2003 Juli 4;330(2):343–58.

217. Nikaido K, Liu PQ, Ames GF. Purification and characterization of HisP, the ATP-binding subunit of a traffic ATPase (ABC transporter), the histidine permease of Salmonella typhimurium. Solubility, dimerization, and ATPase activity. J. Biol. Chem. 1997 Okt 31;272(44):27745–52.

218. Smith PC, Karpowich N, Millen L, Moody JE, Rosen J, Thomas PJ, u. a. ATP binding to the motor domain from an ABC transporter drives formation of a nucleotide sandwich dimer. Mol. Cell. 2002 Juli;10(1):139–49.

219. Senior AE, al-Shawi MK, Urbatsch IL. The catalytic cycle of P-glycoprotein. FEBS Lett. 1995 Dez 27;377(3):285–9.

220. Payen LF, Gao M, Westlake CJ, Cole SPC, Deeley RG. Role of carboxylate residues adjacent to the conserved core Walker B motifs in the catalytic cycle of multidrug resistance protein 1 (ABCC1). J. Biol. Chem. 2003 Okt 3;278(40):38537–47.

221. Martin C, Berridge G, Mistry P, Higgins C, Charlton P, Callaghan R. Drug binding sites on P-glycoprotein are altered by ATP binding prior to nucleotide hydrolysis. Biochemistry. 2000 Okt 3;39(39):11901–6.

222. Rothnie A, Theron D, Soceneantu L, Martin C, Traikia M, Berridge G, u. a. The importance of cholesterol in maintenance of P-glycoprotein activity and its membrane perturbing influence. Eur. Biophys. J. 2001 Okt;30(6):430–42.

223. Friche E, Demant EJ, Sehested M, Nissen NI. Effect of anthracycline analogs on photolabelling of p-glycoprotein by [125I]iodomycin and [3H]azidopine: relation to lipophilicity and inhibition of daunorubicin transport in multidrug resistant cells. Br. J. Cancer. 1993 Feb;67(2):226–31.

224. Dawson RJ, Locher KP. Structure of a bacterial multidrug ABC transporter. Nature. 2006 Sep;443:180–5.

225. Ward A, Reyes CL, Yu J, Roth CB, Chang G. Flexibility in the ABC transporter MsbA: Alternating access with a twist. Proc. Natl. Acad. Sci. U.S.A. 2007 Nov;104:19005–10.

226. Ravna AW, Sylte I. Homology modeling of transporter proteins (carriers and ion channels). Methods Mol. Biol. 2012;857:281–99.

227. Parveen Z, Stockner T, Bentele C, Pferschy S, Kraupp M, Freissmuth M, u. a. Molecular dissection of dual pseudosymmetric solute translocation pathways in human p-glycoprotein. Mol. Pharmacol. 2011 März;79:443–52.

228. Ma Q, Lu AYH. Pharmacogenetics, pharmacogenomics, and individualized medicine. Pharmacol. Rev. 2011 Juni;63(2):437–59.

229. Türk D, Szakács G. Relevance of multidrug resistance in the age of targeted therapy. Curr Opin Drug Discov Devel. 2009 März;12(2):246–52.

230. Türk D, Hall MD, Chu BF, Ludwig JA, Fales HM, Gottesman MM, u. a. Identification of compounds selectively killing multidrug-resistant cancer cells. Cancer Res. 2009 Nov 1;69(21):8293–301.

231. Hall MD, Handley MD, Gottesman MM. Is resistance useless? Multidrug resistance and collateral sensitivity. Trends Pharmacol. Sci. 2009 Okt;30(10):546–56.

232. Hall MD, Brimacombe KR, Varonka MS, Pluchino KM, Monda JK, Li J, u. a. Synthesis and structure-activity evaluation of isatin-β-thiosemicarbazones with improved selective activity toward multidrug-resistant cells expressing P-glycoprotein. J. Med. Chem. 2011 Aug 25;54(16):5878–89.

233. Bell SE, Quinn DM, Kellett GL, Warr JR. 2-Deoxy-D-glucose preferentially kills multidrug-resistant human KB carcinoma cell lines by apoptosis. Br. J. Cancer. 1998 Dez;78(11):1464–70.

234. Warr JR, Bamford A, Quinn DM. The preferential induction of apoptosis in multidrug-resistant KB cells by 5-fluorouracil. Cancer Lett. 2002 Jan 10;175(1):39–44.

235. Warr JR, Brewer F, Anderson M, Fergusson J. Verapamil hypersensitivity of vincristine resistant Chinese hamster ovary cell lines. Cell Biol. Int. Rep. 1986 Mai;10(5):389–99.

236. Warr JR, Anderson M, Fergusson J. Properties of verapamil-hypersensitive multidrug-resistant Chinese hamster ovary cells. Cancer Res. 1988 Aug 15;48(16):4477–83.

237. Cano-Gauci DF, Riordan JR. Action of calcium antagonists on multidrug resistant cells. Specific cytotoxicity independent of increased cancer drug accumulation. Biochem. Pharmacol. 1987 Juli 1;36(13):2115–23.

238. Tiberghien F, Loor F. Ranking of P-glycoprotein substrates and inhibitors by a calcein-AM fluorometry screening assay. Anticancer Drugs. 1996 Juli;7:568–78.

239. Litman T, Zeuthen T, Skovsgaard T, Stein WD. Structure-activity relationships of P-glycoprotein interacting drugs: kinetic characterization of their effects on ATPase activity. Biochim. Biophys. Acta. 1997 Aug 22;1361(2):159–68.

240. Pauli-Magnus C, von Richter O, Burk O, Ziegler A, Mettang T, Eichelbaum M, u. a. Characterization of the major metabolites of verapamil as substrates and inhibitors of P-glycoprotein. J. Pharmacol. Exp. Ther. 2000 Mai;293(2):376–82.

241. Shi JG, Zhang Y, Yeleswaram S. The relevance of assessment of intestinal P-gp inhibition using digoxin as an in vivo probe substrate. Nat Rev Drug Discov. 2011 Jan;10(1):75; author reply 75.

242. Szakacs G, Annereau JP, Lababidi S, Shankavaram U, Arciello A, Bussey KJ, u. a. Predicting drug sensitivity and resistance: profiling ABC transporter genes in cancer cells. Cancer Cell. 2004 Aug;6:129–37.

243. Rydzewski RM. Real World Drug Discovery - A Chemist`s Guide to Biotech and Pharmaceutical Research. 1. Aufl. Elsevier; 2008.

244. Lombardino JG, Lowe JA 3rd. The role of the medicinal chemist in drug discovery--then and now. Nat Rev Drug Discov. 2004 Okt;3(10):853–62.

245. Ripphausen P, Stumpfe D, Bajorath J. Analysis of structure-based virtual screening studies and characterization of identified active compounds. Future Med Chem. 2012 Apr;4(5):603–13.

246. Kitchen DB, Decornez H, Furr JR, Bajorath J. Docking and scoring in virtual screening for drug discovery: methods and applications. Nat Rev Drug Discov. 2004 Nov;3(11):935–49.

247. Wasserman SR, Koss JW, Sojitra ST, Morisco LL, Burley SK. Rapid-access, high-throughput synchrotron crystallography for drug discovery. Trends Pharmacol. Sci. 2012 Mai;33(5):261–7.

248. Blundell TL, Jhoti H, Abell C. High-throughput crystallography for lead discovery in drug design. Nat Rev Drug Discov. 2002 Jan;1(1):45–54.

249. Nurisso A, Daina A, Walker RC. A practical introduction to molecular dynamics simulations: applications to homology modeling. Methods Mol. Biol. 2012;857:137–73.

250. Bordner AJ. Force fields for homology modeling. Methods Mol. Biol. 2012;857:83–106.

251. Venclovas C. Methods for sequence-structure alignment. Methods Mol. Biol. 2012;857:55–82.

252. Liu T, Tang GW, Capriotti E. Comparative modeling: the state of the art and protein drug target structure prediction. Comb. Chem. High Throughput Screen. 2011 Juli;14(6):532–47.

253. Tanrikulu Y, Schneider G. Pseudoreceptor models in drug design: bridging ligand- and receptor-based virtual screening. Nat Rev Drug Discov. 2008 Aug;7(8):667–77.

254. Mannhold R, Krogsgaard-Larsen P, Timmermann H. QSAR: Hansch Analysis and Related Approaches. Weinheim, Germany: WILEY-VCH; 1993.

255. Hansch C, Leo A. Exploring QSAR, Fundamentals and Applications in Chemistry and Biology. Washington, DC: American Chemical Society; 1995.

256. Langer T, Hoffmann RD. Pharmacophores and Pharmacophore Searches. Weinheim, Germany: WILEY-VCH; 2006.

257. Willett P. Similarity searching using 2D structural fingerprints. Methods Mol. Biol. 2011;672:133–58.

258. Mitchell TM. Machine Learning. McGraw-Hill; 1997.

259. Mitchell JBO. Informatics, machine learning and computational medicinal chemistry. Future Med Chem. 2011 März;3(4):451–67.

260. Gedeck P, Kramer C, Ertl P. Computational analysis of structure-activity relationships. Prog Med Chem. 2010;49:113–60.

261. Todeschini R, Consonni V. Handbook of Molecular Descriptors. Weinheim, Germany: WILEY-VCH; 2000.

262. Witten I, Frank E. Data Mining: Practical Machine Learning Tools and Techniques. Morgan-Kaufman; 2005.

263. Janecek AG., Gansterer W, Demel MA, Ecker GF. On the Relationship Between Feature Selection and Classification Accuracy. JMLR Workshop and Conference Proceedings. 2008. p. 90–105.

264. Dietterich TG. Ensemble Methods in Machine Learning. Multiple Classifier Systems: Lecture Notes in Computer Science. 2000;1857(2000):1–15.

265. Friedman J, Hastie T, Tibshirani R. The Elements of Statistical Learning - Data Mining, Inference and Prediction. 2nd. Ed. Stanford, California; 2008.

266. Sakiyama Y. The use of machine learning and nonlinear statistical tools for ADME prediction. Expert Opin Drug Metab Toxicol. 2009 Feb;5(2):149–69.

267. Jolliffe IT. Principal Component Analysis. 2nd Ed. Springer; 2002.

268. Cadima J, Jolliffe IT. Variable Selection and the Interpretation of Principal Subspaces. Journal of Agricultural, Biological and Environmental Statistics. 2001 März;6:62–79.

269. Cadima J, Jolliffe IT. Loading and correlations in the interpretation of principle compenents. Journal of Applied Statistics. 1995;1995(22):2.

270. Cadima J, Orestes Cerdeira J, Minhoto M. Computational aspects of algorithms for variable selection in the context of principal components. Computational Statistics and Data Analysis. 2004 Sep;47:225–36.

271. Kohavi JG, Pflegger K. Irrelevant features and the subset selection problem. Machine Learning: Proceedings of the Eleventh International Conference. 1994;121–9.

272. Yu L, Liu H. Efficient feature selection via analysis of relevance and redundancy. Journal of Machine Learning Research. 2004;1205–24.

273. Hall MA. PhD Thesis: Correlation-based Feature Selection for Machine Learning. 1999.

274. Guyon I, Elisseeff A. An introduction to variable and feature selection. Journal of Machine Learning Research. 2003;3:1157–82.

275. Golub TR, Slonim DK, Huard C, Gaasenbeek M, Mesirov JP, Lander ES. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science. 1999;286(5439):531–7.

276. Guyon I, Gunn S, Nikravesh M, Zadeh L. Feature Extraction, Foundations and Applications. Physica-Verlag, Springer; 2006.

277. Kira K, Rendell LA. A practical approach to feature selection. Proceedings of the ninth international workshop on Machine learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 1992. p. 249–56.

278. Kononenko I. Estimating Attributes: Analysis and Extensions of RELIEF. In: Bergadano F, Raedt LD, Herausgeber. European Conference on Machine Learning. Springer; 1994. p. 171–82.

279. Fix E, Hodges JL. Discriminatory analysis, nonparametric discrimination: Consistency properties. USAF School of Aviation Medicine; 1951.

280. Shen M, LeTiran A, Xiao Y, Golbraikh A, Kohn H, Tropsha A. Quantitative structure-activity relationship analysis of functionalized amino acid anticonvulsant agents using k nearest neighbor and simulated annealing PLS methods. J. Med. Chem. 2002 Juni;45:2811–23.

281. Cedeno W, Agrafiotis DK. Using particle swarms for the development of QSAR models based on K-nearest neighbor and kernel regression. J. Comput. Aided Mol. Des. 2003;17:255–63.

282. Ajmani S, Jadhav K, Kulkarni SA. Three-dimensional QSAR using the k-nearest neighbor method and its interpretation. J Chem Inf Model. 2006;46:24–31.

283. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, u. a. Enhancing the effectiveness of similarity-based virtual screening using nearest-neighbor information. J. Med. Chem. 2005 Nov;48:7049–54.

284. Wang XS, Tang H, Golbraikh A, Tropsha A. Combinatorial QSAR modeling of specificity and subtype selectivity of ligands binding to serotonin receptors 5HT1E and 5HT1F. J Chem Inf Model. 2008 Mai;48:997–1013.

285. Vasanthanathan P, Taboureau O, Oostenbrink C, Vermeulen NP, Olsen L, J?rgensen FS. Classification of cytochrome P450 1A2 inhibitors and noninhibitors by machine learning techniques. Drug Metab. Dispos. 2009 März;37:658–64.

286. Suenderhauf C, Hammann F, Maunz A, Helma C, Huwyler J. Combinatorial QSAR modeling of human intestinal absorption. Mol. Pharm. 2011 Feb;8:213–24.

287. Gupta SP, Samanta S, Patil VM. A 3D-QSAR study on a series of benzimidazole derivatives acting as hepatitis C virus inhibitors: application of kNN-molecular field analysis. Med Chem. 2010 März;6:87–90.

288. Demel M, Janecek AG., Gansterer W, Ecker GF. Comparison of Contemporary Feature Selection Algorithms: Application to the Classification of ABC-Transporter Substrates. QSAR & Comb. Sci. (10):1087–91.

289. Murthy. Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey. Data Mining and Knowledge Discovery 2. 1998;2:345–89.

290. Quinlan JR. Induction of Decision Trees. Mach. Learn. 1986;1:81–106.

291. Quinlan JR. C4.5: Programs for Machine Learning. Morgan Kaufman Publishers; 1993.

292. Quinlan JR. Improved use of continuous attributes in c4.5. Journal of Artificial Intelligence Research. 1996;4:77–90.

293. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Wadsworth & Brooks/Cole Advanced Books & Software; 1984.

294. Breiman L. Random Forests. Machine Learning. 2001;

295. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random Forests: a classification and regression tool for compound classification and QSAR modeling. J Chem Inf Comput Sci. 2003 Dez;43(6):1947–58.

296. Segal MR. Machine Learning Benchmarks and Random Forests Regression [eScholarship] [Internet]. 2004 [zitiert 2012 Aug 21]. Available von: http://escholarship.org/uc/item/35x3v9t4#page-1

297. Deng H, Runger G, Tuv E. Bias of importance measures for multi-valued attributes and solutions. Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN). 2011;293–300.

298. Liaw A, Wiener M. Classification and Regression by randomForests. R News. 2002;2(3):18–22.

299. Friedman J, Popescu B. Predictive Learning via Rule Ensembles. The Annals of Applied Statistics. 2008;2(3):916–54.

300. Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995;20(3):273–97.

301. Suykens JAK. Support vector machines: A nonlinear modelling and control perspective. Eur. J. Control. 2001;7:311–27.

302. Campbell C. Kernel methods: a survey of current techniques. Neurocomputing. 2002;48:63–84.

303. Ivanciuc Q. Applications of Support Vector Machines in Chemistry. Reviews in Computational Chemistry. Weinheim: Wiley-VCH; 2007. p. 291–400.

304. Golbraikh A, Shen M, Xiao Z, Xiao Y-D, Lee K-H, Tropsha A. Rational selection of training and test sets for the development of validated QSAR models. J. Comput. Aided Mol. Des. 2003 Apr;17(2-4):241–53.

305. Golbraikh A, Tropsha A. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. J. Comput. Aided Mol. Des. 2002 Juni;16(5-6):357–69.

306. Willett P. Dissimilarity-Based Algorithms for Selecting Structurally Diverse Sets of Compounds. Journal of Computational Biology. 1999;4467–457.

307. Chirico N, Gramatica P. Real external predictivity of QSAR models: how to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. J Chem Inf Model. 2011 Sep 26;51(9):2320–35.

308. Chirico N, Gramatica P. Real External Predictivity of QSAR Models. Part 2. New Intercomparable Thresholds for Different Validation Criteria and the Need for Scatter Plot Inspection. Journal of chemical information and modeling [Internet]. 2012 Juli 13 [zitiert 2012 Aug 23]; Available von: http://www.ncbi.nlm.nih.gov/pubmed/22721530

309. Baldi P, Brunak S, Chauvin Y, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinfomatics Review. 2000;16(5):412–24.

310. Wess G. How to escape the bottleneck of medicinal chemistry. Drug Discov. Today. 2002 Mai 15;7(10):533–5.

311. Guha R, Jurs PC. Determining the validity of a QSAR model--a classification approach. J Chem Inf Model. 2005 Feb;45(1):65–73.

312. Validation of (Q)SAR Models [Internet]. [zitiert 2012 Feb 21]. Available von: http://www.oecd.org/document/4/0,3746,en_2649_34379_42926724_1_1_1_1,00.html

313. Netzeva TI, Aptula AO, Benfenati E, Cronin MTD, Gini G, Lessigiarska I, u. a. Description of the electronic structure of organic chemicals using semiempirical and ab initio methods for development of toxicological QSARs. J Chem Inf Model. 2005 Feb;45(1):106–14.

314. Sushko I, Novotarskyi S, Körner R, Pandey AK, Cherkasov A, Li J, u. a. Applicability domains for classification problems: Benchmarking of distance to models for Ames mutagenicity set. J Chem Inf Model. 2010 Dez 27;50(12):2094–111.

315. Tetko IV, Sushko I, Pandey AK, Zhu H, Tropsha A, Papa E, u. a. Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection. J Chem Inf Model. 2008 Sep;48(9):1733–46.

316. Zhu H, Tropsha A, Fourches D, Varnek A, Papa E, Gramatica P, u. a. Combinatorial QSAR modeling of chemical toxicants tested against Tetrahymena pyriformis. J Chem Inf Model. 2008 Apr;48:766–84.

317. Tropsha A, Golbraikh A. Predictive QSAR modeling workflow, model applicability domains, and virtual screening. Curr. Pharm. Des. 2007;13(34):3494–504.

318. Jaworska J, Nikolova-Jeliazkova N, Aldenberg T. QSAR applicabilty domain estimation by projection of the training set descriptor space: a review. Altern Lab Anim. 2005 Okt;33(5):445–59.

319. Tetko IV, Bruneau P, Mewes H-W, Rohrer DC, Poda GI. Can we estimate the accuracy of ADME-Tox predictions? Drug Discov. Today. 2006 Aug;11(15-16):700–7.

320. Johnson M, Maggiora GM. Concepts and Applications of Molecular Similarity. New york: John Wiley and Sons; 1990.

321. Peltason L, Bajorath J. Molecular Similarity Analysis Uncovers Heterogeneous Structure-Activity Relationships and Variable Activity Landscapes. Chemical Biology. 2007;14:489–97.

322. Maggiora GM. On Outliers and Activity Cliffs. Why QSAR Often Disappoints. J. Chem. Inf. Mod. 2006;46:1535–1535.

323. Eckert H, Bajorath J. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. Drug Discov. Today. 2007 März;12(5-6):225–33.

324. Bajorath J, Peltason L, Wawer M, Guha R, Lajiness MS, Van Drie JH. Navigating structure-activity landscapes. Drug Discov. Today. 2009 Juli;14(13-14):698–705.

325. Peltason L, Bajorath J. Quantifying the Nature of Structure−Activity Relationships. Journal of Medicnal Chemistry. 2007;50:5571–8.

326. Guha R, Van Drie JH. Structure-Activity Landscape Index: identifying and quantifying activity cliffs. J. Chem. Inf. Mod. 2008;48:646–58.

327. Mestres J, Martin-Couce L, Gregori-Puigjane E, Cases M, Boyer S. Ligand-Based Approaches to in Silico Pharmacology: Nuclear Receptor Profiling. J. Chem. Inf. Mod. 2006;2725–36.

328. Paolini GV, Shapland RHB, van Hoorn WP, Mason JS, Hopkins AL. Global mapping of pharmacological space. Nat. Biotechnol. 2006 Juli;24(7):805–15.

329. Hert J, Keiser MJ, Irwin JJ, Oprea TI, Shoichet BK. Quantifying the relationships among drug classes. J Chem Inf Model. 2008 Apr;48(4):755–65.

330. Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, u. a. Predicting new molecular targets for known drugs. Nature. 2009 Nov 12;462(7270):175–81.

PART C: OBJECTIVES & AIMS

Human ABC-transporters, which act as drug carriers, are notorious for their pivotal role in influencing the pharmacokinetic fate of a plethora of marketed drugs and also for their contribution to MDR, a leading cause of failure of anti-cancer pharmacotherapy in clinical practice. *In-silico* methods have gained a lot of acceptance in the last years with respect to understand the molecular triggers that drive biological activity of small molecules on the one hand but also with respect to support rational decision making in early phases of drug development on the other hand. However, some of the models for molecules interacting with humanABC-transporter published in the literature lack predictive accuracy but are easy to interpret, whereas others are highly accurate but provide only limited information on a chemical basis.

This thesis aims to apply ligand-based *in-silico* machine learning methods in order to provide useful insights into the underlying medicinal chemistry of compounds interacting with ABC-transporters with a special **focus on ABCB1**, the paradigm transporter.

Primary Objective

- The primary objective of this thesis is to provide ligand-based machine learning models for the **ABCB1 substrate/non-substrate** classification problem (using data sets retrieved from different sources) that are both; highly **predictive** but at the same time are also easy to **interpret.**

Secondary Objectives

In order to accomplish this primary objective, different methodological approaches that encompass all different steps of a typical machine learning workflow will be explored:

- Evaluation of **data pre-processing** methods.
    - An important step in machine learning is the selection of an optimal model out of a set of plausible models. A promising strategy is to perform **feature selection**, which aims to retain only those structural or physico-chemical features (descriptors) that are most suitable for modelling purposes. However, a variety of feature selection algorithms exists and they are based on different mathematical definitions of feature relevance and therefore can provide varying results. The comparative assessment of different feature selection algorithms is a particular part of this thesis.
    - Another way to provide useful models that might be easy to interpret is to **develop new descriptors** that describe the chemical information of the data. Therefore, a modification of the previously introduced *Si*milarity-*b*ased S*AR* (SIBAR) concept that utilizes similarity values to characterize molecules instead of structural or physico-chemical properties will be presented and critically appraised with respect to its predictive performance and interpretability.
- Application of **ensemble algorithms**.
    - One of the most promising and also novel strategies to obtain models that show high predictive performance is to construct an ensemble of models. These ensemble methods have the advantage to outperform so-called single

classifiers, but often lack the opportunity to provide informative insights into the underlying properties of the data set under consideration ("black box"-models that are not interpretable) due to their high complexity. Herein, promising and contemporary ensemble algorithms will be applied that have been developed to overcome this disadvantage. The interpretation of these rather complex models is a central objective of this thesis.

- Evaluation of **post-processing** methods.
    - Recently, the importance of **assessing the applicability domain** of an *in-silico* machine learning model with respect to its performance when applied in practical, "real-life" situations has been recognized and can be regarded as probably one of the most progressing fields in contemporary pharmacoinformatics. In order to gain insights into the generalizability of ABCB1 substrate/non-substrate classification models, different measures of the model`s applicability domain will be applied and carefully evaluated with respect to their feasibility.

Tertiary Objectives

- An additional objective of this thesis is to characterize **MDR-selective compounds** using *in-silico* methods. MDR-selective compounds constitute a novel pharmacological class that might represent innovative chemotherapeutics for future cancer therapy owing to their potential to selectively kill MDR cells. Here, the focus is also set on the the interpretation of the generated models with the aim to elucidate and characterize the SAR of these molecules.

PART D: PUBLICATIONS & RESULTS

Chapter 9: Publication I - In silico prediction of substrate properties for ABC-multi-drug transporters

Abstract

Overexpression of ABC (ATP-binding cassette)-type drug efflux pumps, such as ABCB1, ABCC1 and ABCG2 in cancer cells confers multi-drug resistance (MDR) and represents a major cause of treatment failures in cancer therapy. Furthermore, there is increasing evidence for the important contribution of ABC-transporters to bioavailability, distribution, elimination and blood–brain barrier permeation of drug candidates. This review presents an overview on the different computational methods and models pursued to predict ABC-transporter substrate properties of drug-like compounds. They range from linear discriminant analysis to pharmacophore modelling and machine learning algorithms. Many of these models show a satisfying performance within the study-specific, defined chemical space but general applicability for the whole drug-like chemical space still needs to be proven. First attempts aiming towards selectivity profiling for ligands of the two polyspecific transporters ABCB1 and ABCG2 is also discussed. This might pave the way for a pharmacological profiling of compound series with special focus on their ADMET (absorption, distribution, metabolism, excretion and toxicity) properties.

Introduction

The ABC (ATP-binding cassette)-transporter superfamily consists of seven families (A – G, based on their sequence homology) and contains 48 proteins [1]. Several members of this family, such as ABCB1 (P-glycoprotein, P-gp), ABCC1 (multi-drug resistance protein 1, MRP1) and ABCG2 (breast cancer related protein, BCRP) are strongly connected to multi-drug resistance (MDR) in cancer therapy [2-4]. They act as exporter of a broad variety of xenobiotics thus decreasing intracellular accumulation of anticancer drugs. However, besides their role in tumour resistance, several ABC-transporters have also been shown to be involved in bioavailability, distribution and elimination of drugs. Furthermore, human ABC-transporters that do not participate in drug efflux, such as cystic fibrosis conductance regulator, ABCC7 and the SURs (sulphonylurea receptors, ABCC8, ABCC9) represent ion channels that have been linked to severe diseases. In the further context of this article we refer to ABC multi-drug (xenobiotic) transporters. ABC multi-drug transporters are constitutively expressed in many organs and are found especially in blood–tissue barriers such as the blood–brain barrier, the blood–testis barrier, the bile canalicular membrane of hepatocytes, the intestine and the kidney [5,6]. Bioavailability of drug candidates as well as the probability of drug/drug interactions is therefore strongly influenced by their ability to interact with ABC-transporters. Subsequent to the pioneering experiments by Schinkel *et al.* [7] with mdr1a double knockout mice numerous reports appeared in the literature showing the importance of drug/transporter interaction for bioavailability, disposition and brain uptake [8]. Furthermore, also nutrients such as grapefruit juice are increasingly recognised as being responsible for altered bioavailability and metabolism of drugs through their interaction with ABCB1 and MRP2 [9]. This prompted Wu and Benet [10] to propose a biopharmaceutical drug disposition classification system that considers the influence of both uptake and efflux transporters. Finally, there is increasing evidence that cholestatic forms of drug-induced liver damage

result from a drug- or metabolite-mediated inhibition of hepatobiliary transporter systems, such as ABCB1, ABCB4, ABCG2, ABCG5 and ABCG8 [11]. Initially, ABC-transporters were considered as versatile targets to overcome MDR in tumour therapy and research therefore concentrated on the development of inhibitors [12]. However, clinical studies of several compounds revealed considerable side effects and no compound has been approved yet. Within the past decade the focus of pharmaceutical research moved more from inhibitor design towards lowering the likelihood that a new compound will be a transporter substrate. As ABC-transporters are not only responsible for cancer cell resistance to cytostatic drugs but also play an important role concerning drug–drug interactions, bioavailability and clearance, the prediction of substrate properties gains more and more significance. This review outlines the current status of *in silico* models for prediction of substrates of ABC-efflux pumps mainly focusing on ABCB1 (P-glycoprotein), the paradigm transporter in the field.

Biological Assays for ABC-transporter substrate properties

Accurate biological assays are the basis for generation of *in silico* models, as the output of these experiments derives the Y-(response) variable in a QSAR (quantitative structure–activity relationships) or classification study. Therefore, the different approaches used for assessing ABC-transporter substrate properties is discussed briefly. In contrast to classical inhibition assays, the measurement of transporter substrate properties is more difficult, as five possibilities have to be considered:

- Compounds are extracted from the membrane and transported to the outer medium faster than rediffusion can occur. The pump is able to establish a concentration gradient and the compound is rapidly extracted from the cell. Such a compound is defined as a substrate (**Figure 1A**).
- Compounds bind to the transporter and inhibit its activity. No concentration gradient is established (**Figure 1B**).
- Compounds are extracted from the membrane and transported to the outer medium but rediffuse equally rapidly. The pump is only able to establish a small concentration gradient at utmost. In inhibition assays such a compound acts like an inhibitor. This type of behaviour is often referred to as modulator (**Figure 1C**).
- Compounds diffuse through the membrane and do not interact with the respective transporter (**Figure 1D**).
- Compounds passively diffuse back with a rate comparable to its active efflux. As a result a measurable concentration gradient cannot be maintained (**Figure 1E**).

The scenarios described above clearly highlight that only a combination of assays can describe the full interaction spectrum of a given compound with a particular transporter. This is a fundamental reason why different studies assign the same compound differently, which additionally exacerbates literature mining. *In vivo* assays mostly rely on the generation of transgenic or mutant animals and provide insight on the impact of ABC-transporters on the pharmacokinetic profiles of drug candidates. A standard approach is performed by assessing the brain/plasma exposure ratios in mdr1a/mdr1b knockout mice versus wild-type mice [13]. For a substrate of the ABC-transporter ABCB1 the concentration ratio brain to blood is

considerably enhanced in transporter knockout animals when compared to the wild-type animals. However, although reflecting best the real life situation, this method is not very useful as input for computational studies, as it comprises the sum of several effects such as solubility, plasma protein binding, passive diffusion and active transport. The *in vitro* assays can be divided into six different classes:

- assays that monitor the ATPase activity of ABC transporters;
- bidirectional transport assays on confluent cell monolayers (e.g., monolayer efflux assay);
- assays that look for competitive substrates;
- toxicity assays that compare transporter overexpressing and wild-type cell lines;
- correlating transporter expression to compound toxicity over a panel of different, well-characterised tumour cell lines; and
- induction of the transporter.

1. As ABC-transporters harbour an intrinsic ATPase activity (contrary to other ATPases, e.g., Na/K-ATPase), the ATPase activity assay mainly aims to report changes in the basal ATP hydrolysis rate in the presence of compounds of interest [14]. The modulation of ATPase activity by a compound definitely is a strong hint for interaction with the transporter. However, in recent studies that aimed to compare several assay strategies, it was shown that certain compounds (e.g., cyclosporin A) gave a positive readout in an efflux assay but were not identified as substrates in ATPase assays [15]. Furthermore, we experienced that some propafenones stimulate ATPase activity of ABCB1 in low concentrations and inhibit it in high concentrations, some show a pure stimulation and some show only inhibition [16]. These findings show that conclusions drawn solely from the ATPase assay need to be taken cautiously.

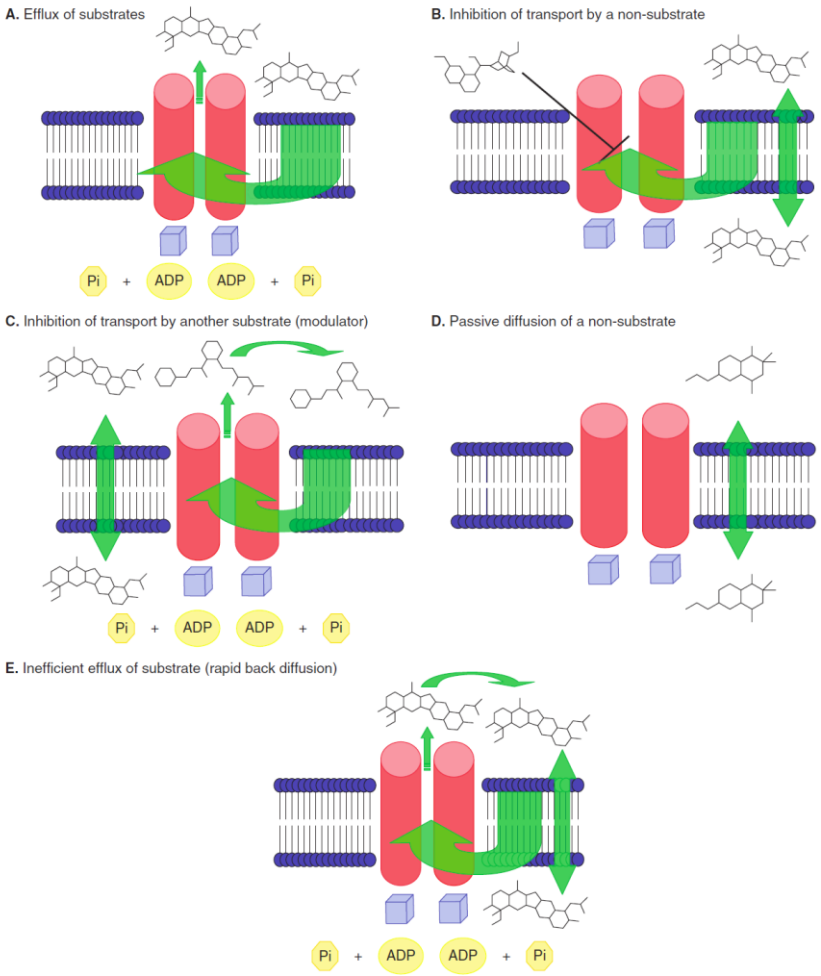


Figure 1: Substrates/inhibitors/modulators/non-substrates interacting with ABC-transporters.

2. Several studies showed that the direct measurement of transport across adherent cell monolayers is the most reliable method for the identification of ABC-transporter substrates [15,17]. The permeation ratio of a compound basolateral-to-apical (B → A) and apical-to-basolateral (A → B) is compared to the respective ratio in the presence of a transporter inhibitor. The monolayer efflux assay is labour intensive and of course dependent on the passive permeability of the test compound. Furthermore, for highly hydrophobic compounds the unstirred water layer near the cell surface may become rate limiting. This assay also allows testing compounds for being modulators or inhibitors by evaluating their capacity to alter the transepithelial flux of a reference substrate. This allows the differentiation between compounds acting as substrates and modulators or inhibitors. Furthermore, it is also possible to discriminate between modulator and inhibitor by testing it directly as being transported or not.
3. Evaluation of transport activity of ABCB1 can also be performed by the fluorometric measurement of the intracellular accumulation of calcein or a different flurophore [18]. The assay is based on the fact that ABCB1 can transport calcein–acetoxy–methylester (calcein–AM) from cells but not free calcein. Calcein–AM enters cells by passive diffusion and is effluxed by ABCB1 before it is hydrolysed to calcein by intracellular esterases. Compounds that compete with calcein–AM prevent its extrusion and intracellular fluorescence increases in a time-dependent manner. A

disadvantage of this assay is that it cannot distinguish between modulators and inhibitors, as both types of compounds will show a similar experimental readout.

4. The sensitivity of a transporter overexpressing cell line to a cytotoxic compound compared to the sensitivity of the wild-type cell line is another way to assess whether the compound is likely to be a substrate of that transporter. Zhou *et al.* [19] measured activities against the paclitaxel resistant and ABCB1-overexpressing cell line H460 taxR to design compounds that are not substrates of ABCB1 and are effective against drug-resistant cancer cell lines. The major disadvantage of this assay is that it is only applicable to cytotoxic compounds.
5. A different approach to predict drug sensitivity and resistance of ABC-transporters was conducted by the group of Gottesman [20]. They used quantitative real-time RT–PCR (polymerase chain reaction after reverse transcription of RNA) to profile the mRNA expression level of the 48 human ABC-transporters in 60 different human cancer cell lines to screen for anticancer activity of 1429 drugs or drug candidates. The transcript expression level of each individual transporter was correlated with the growth inhibitory effect of each compound under investigation across the 60 cell lines. According to this method a compound's liability for a particular transporter is given in form of the Pearson's correlation coefficient. A negative correlation coefficient suggests that it is a substrate for a particular transporter (high transporter expression level, low cytotoxicity of the compound) and a non-correlation suggests that a substance is not interacting with the transporter. To confirm that this procedure is really able to identify new substrates for ABCB1, they additionally apply the MTT (3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium-bromide) assay using human carcinoma cell lines that show a high ABCB1 expression level. The impact of this approach is evident as it allocates biological activity for 1429 compounds for all human ABC-transporters. A potential drawback of this screening methodology is that the provided data are extremely 'noisy' because it is based on the correlation of 60 different cell lines.
6. A rather global approach is to identify substrates or modulators of ABC-transporters through their influence on the expression level of the protein [21-23] . The underlying rationale of this approach is that xenobiotics (i.e., in this case: substrates/modulators) induce cellular defence mechanisms. However, it is known that not every ligand necessarily modulates the expression level of either of the transporters considered. Additionally this approach again is not able to differentiate between substrates and modulators. Therefore, the relevance of this assay seems to be questionable.

As the *in vivo* and *in vitro* assays are time-consuming and cannot be standardised suitable for a high-throughput format, they are used for small sets of compounds. This is reflected by the limited amount of data available in the literature that can serve as training and test sets for generation of *in silico* models. Furthermore, almost 20% of the compounds published are inconsistently assigned. This is most probably owing to different interpretation of assay results. Especially the group of modulators requires a careful consideration, as in the various

assays described they can show both inhibitor- and substrate-like behaviour (e.g., verapamil, cyclosporin, propafenone).

Computational studies for the prediction of ABC-transporter substrate properties

Pharmacophore models for ABC-transporter substrates

The underlying hallmark of a pharmacophore model relies on the assumption that structurally diverse molecules establish a similar interaction pattern when binding to a receptor [24]. Thus, pharmacophore models direct the medicinal chemist in illuminating the key ligand–receptor interactions without knowing the structure of the receptor. A recent up-to-date overview of the tools used at present in this research area has been provided by Wolber and colleagues [25]. Additionally, Chang et al. [26] provide a comprehensive review on pharmacophore modelling studies on a variety of drug transporters. In the following section we focus on the current progress of pharmacophore models of ABCB1 substrates. The foundation of the ABCB1 pharmacophore was laid with the structure–function analyses of reserpine and yohimbine analogues in a human leukaemia cell line [27,28]. Two aromatic features as well as the presence of basic nitrogen were proposed as necessary elements of these ABCB1 ligands. The importance of a nitrogen atom has also been highlighted by other authors and it was always interpreted as positive ionisable group. However, using a series of propafenone analogues systematically varied in the vicinity of the nitrogen atom we could demonstrate that the nitrogen atom interacts as hydrogen-bond acceptor rather than as positively charged entity [29]. An assessment of 3D structures proposed by Seelig showed a general pattern for ABCB1 substrate recognition consisting of two- or three electron donor groups with a fixed spatial separation of 2.5 ± 0.3 Å (type I pattern) or 4.6 ± 0.6 Å (type II pattern), respectively [30] . Notably, this study does not only discriminate between substrates and non-substrates but also differentiates between MDR-inducers and borderline substrates. Borderline substrates are defined as substrates with a very low substrate activity. Penzotti proposed an ensemble model of 100 pharmacophores, consisting of a set of two-, three- and four-point pharmacophores to distinguish between ABCB1 substrates and non-substrates [31]. The pharmacophore ensemble was derived on the basis of pairwise comparison of both substrates and non-substrates in terms of the relative information content of each pharmacophore. Compounds that fitted at least 20 pharmacophores of the ensemble were assigned to be potential ABCB1 substrates. Most of the highest scoring two-point pharmacophores in this study are in convergence with either the type I- or the type II-ABCB1 recognition patterns mentioned by Seelig [30]. Although this model showed an overall classification accuracy of 80% for the training set, the accuracy for the test set dropped down to 63% (Table 1). Despite several pharmacophore models of ABCB1 inhibitors Ekins et al. [32] also generated a substrate pharmacophore on the basis of verapamil, vinblastine and digoxin. An alignment of these three substrates enabled them to suggest that they share a similar binding site at ABCB1. The pharmacophore comprised multiple hydrophobic and hydrogen bond acceptor features. Additionally, Ekins et al. also propose a high similarity between this pharmacophore and those derived for substrates and inhibitors of the cytochrome-P450 3A enzyme [33,34]. Both proteins prefer ligands with multiple hydrophobic features and at least one hydrogen bond acceptor feature.

Table 1: Summary of published PH4 modelling studies

Study	*N*	*Performance*		*Features*
		Training	*Validation*	
ABCB1 substrates and inhibitors				
Penzotti et al [31]	195	Acc: 80%	Acc: 63%	Ensemble PH4: 53 four-point; 39 three-point; 8 two-point
Ekins et al. [32]	16	r^2=0.96	r^2=0.72 (RG)	2HYD, 1HBA, 1Ar
Pajeva and Wiese [35]	20	n.a.	n.a.	2HYD, 3HBA, 1HBD
	27	r^2=0.77	r^2=0.43	4HYD, 1HBA
Ekins et al. [36]	21	r^2=0.88	r^2=0.31	3Ar, 1HYD
	17	r^2=0.86	r^2=0.64	2HYD, 1HBA, 1HBD
Garrigues et al. [37]	7	n.a.	n.a.	PH1: 1Ar, 2Alkyl, 1e - -don PH2: 1Ar, 3Alkyl, 1e - -don
Langer et al.[39]	15	Class. of 105 propafenones: Acc1:97.2%, Acc0:80%	Screening of the WDI: 28 hits	1HBA, 1HYD, 2Ar, 1PI
Cianchetta et al. [38]	129	r^2=0.81		2HYD, 2HBA, size
Chang et al.[40]	33	R=0.87		4HYD, 1HBA
Li et al. [42]	163	Acc:87.7%		9 models out of 12.6 million four-point PH4s
ABCG2 inhibitors				
Chang et al. [41]	4	n.a.		3HYD, 3HBA
ABCC1 inhibitors				
Chang et al. [41]	5	n.a.		3Ar, 3HBA

The pharmacophore model presented by Pajeva et al. [35], which consists of two hydrophobic points, three hydrogen bond acceptor points and one hydrogen bond donor point, led the authors to generate a hypothesis for the structural variety of ligand recognition by ABCB1: i) the verapamil binding site in ABCB1 harbours points responsible for hydrophobic and hydrogen-bonding interactions; and ii) different drugs are able to interact with different receptor points in different binding modes. At this point it needs to be mentioned that another study by Ekins et al. [32,36] also describes a pharmacophore model for MDR-modulators that act at the verapamil binding site. However, despite some similarities between the two models presented, a direct comparison of the distance matrices of the pharmacophoric features revealed some differences. This reflects the promiscuous ligand interaction pattern observed for ABCB1. Garrigues et al. [37] proposed a model consisting of two pharmacophores that partially overlap. This further supports the hypothesis that there are many overlapping binding sites at ABCB1. In a different study, Cianchetta et al. [38] generate a pharmacophore hypothesis for ABCB1 substrates by using alignment-independent GRIND descriptors on a diverse set of 129 compounds. The final model proposed in this study shows two hydrophobic groups and two hydrogen-bond acceptors. In the field of propafenone-type MDR-modulators we derived a pharmacophore, which was successfully applied to screen the Derwent World Drug Index for ABCB1 ligands [39]. Nine out of 28 hits retrieved have previously been described to act as ABCB1 modulators. In a different publication two inhibitor pharmacophore models and one substrate model were subjected to screen the SCUT database [40]. Also in this case the biological testing of seven selected hits confirmed the plausibility of these pharmacophore models. Modelling studies on other ABC-transporters focused on pharmacophore models for inhibitors. Chang and colleagues [41] recently presented an ABCG2 pharmacophore on the basis of four known ABCG2 inhibitors using the HIPHOP suite in Catalyst. The model consists of three hydrogen-bond acceptors and three hydrophobic features. The pharmacophore has been used for screening a database of 500 marketed drugs. Among the 37 hits, 6 compounds have already been described as ABCG2 inhibitors in the literature. The role of the remaining 31 hits still awaits in vitro testing and confirmation. The same group also established a model for inhibitors of ABCC1 (MRP1). This pharmacophore contains three aromatic and three hydrogen bond acceptor features. Additionally, in silico screening using this model revealed 8 hits out of 500 marketed drugs. Three of these hits are known to reverse ABCC1 associated MDR. Overall, the published pharmacophore models for the different ABC-transporters show some convergence concerning the features responsible for drug–transporter interaction (hydrophobic and hydrogen bond acceptor features). However, most of them are not directly comparable owing to remarkable differences in the distance matrices of the indicated features. This reflects the overlapping substrate specificity of ABC-transporters and suggests that multiple or ensemble pharmacophore models may be a better way to cover the pharmacophore space of ABC-transporters. This has been recently shown by Li et al. [42] who provide an ABCB1 substrate classification study based on many pharmacophores. This approach resulted in an excellent overall prediction accuracy with 87.7% for a training set of 163 compounds and 87.6% for a test set of 97 compounds. Briefly, Li et al. established their multiple pharmacophores in the following way: i) a comprehensive search of all possible pharmacophore configurations (12.6 million) for both substrate and non-

substrate compounds (similar to Penzotti's work [31]); and ii) the generation of a statistical significant ensemble of pharmacophores able to differentiate between ABCB1 substrates and non-substrates. In this step, millions of pharmacophores have been generated and evaluated through a frequency analysis of pharmacophore occurrence and a pharmacophore-specific t -statistic. Consecutively, the nine top-ranked, significant pharmacophores have been selected for the establishment of a classification tree. The classification tree separates compounds into different subsets by grouping together those compounds that map to a specific pharmacophore. Furthermore, each of the nine significant pharmacophores can be used independently. Five of them show accuracy on substrates of 1.00, whereas the worst performing model still shows a value of 0.78. This combination of many pharmacophore models with a classification tree algorithm seems to represent a successful state-of-the-art method to determine substrate properties for ABC-transporters. An extension of this approach to other polyspecific ABC-transporters such as ABCG2 seems to be highly promising. Finally, we could recently demonstrate that ABCB1 and ABCG2, although having overlapping substrate specificity, have defined and distinctly different feature-based interaction characteristics with respect to propafenone analogues [43].

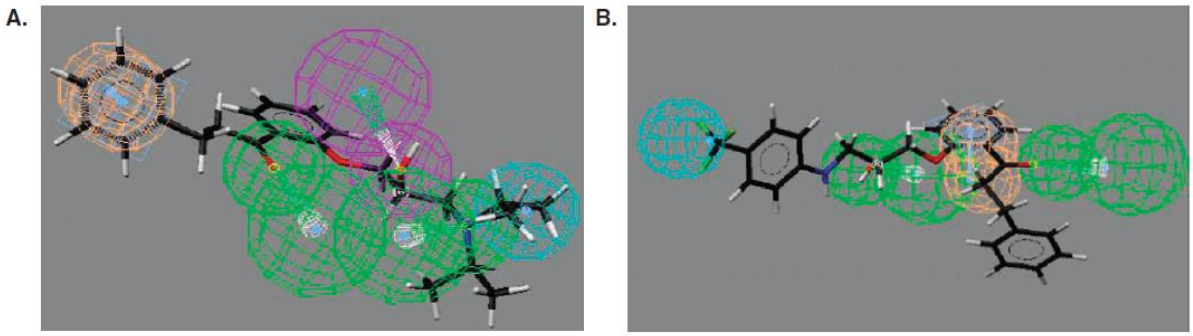


Figure 2: Selectivity profiling of ABC-transporter modulators oft he propafenone-type by pharmacophore modelling.

Classification studies for the prediction of ABC-transporter substrates

In recent years several studies have been conducted to define chemical features commonly shared by ABCB1 substrates by using classification algorithms. Contrary to pharmacophore models, these methods define rules that discriminate substrates and non-substrates on basis of descriptor values reflecting physicochemical properties. Some of them also target the problem of discriminating between substrates and inhibitors [44]. One of the first efforts to define substrate properties for ABCB1 was made by Seelig in 1997 [30,45]. In a study comparing 100 structurally diverse ABCB1 substrates and thereby focusing on the importance of oxygen and nitrogen atoms (= hydrogen-bond acceptors), she concluded that substrate recognition is particularly based on a specific hydrogen-bond pattern. She further showed a correlation between hydrogen-bonding strength and ABCB1 binding strength and defined the partitioning into the membraneas the rate-limiting step (Table 2). The concept of Didziapetris et al. [46] relied on verysimple descriptors and atom counts. Compiling two sets of ABCB1 substrates and non-substrates (220 compounds, 1000 compounds) they postulated the 'rule of four', which states that compounds with the number of hydrogen-bond acceptors in a molecule (N + O) $\geq$ 8, and a molecular mass > 400 Da and most acidic p K a > 4 are likely to be ABCB1 substrates whereas compounds with (N + O) $\leq$ 4, MW < 400 and most basic p K a < 8

probably are nonsubstrates. The respective thresholds were obtained using a sequence of recursive partitioning analyses. Based on these data the authors implemented an ABCB1 substrate specificity prediction tool into their software ADME Boxes [47,48] thereby enabling the user to classify compounds either as ABCB1 substrates, inhibitors or inconclusive. They also provide a reliability index telling the user if compounds lie outside its model applicability domain. For rough classification purposes this tool may give an idea which class one's compounds favour. Similar results have been shown in a recent publication of Gleeson et al. [49] using ~ 2000 compounds. With only four simple ADMET descriptors (molecular mass, log P , positive ionisable and negative ionisable) it was demonstrated that neutral or basic molecules showing a MW > 400 and a logP value > 4 are more likely to be transported by ABCB1 than acidic or zwitterionic compounds. The results are based on a collection of 1975 compounds with measured ABCB1 efflux rations of a proprietary GSK database. In the course of the development of a blood–brain barrier (BBB) permeation model Adenot et al. [50] conducted a study using 91 ABCB1 substrates, taken from Seelig et al. [30]. Using PLS (partial least squares)-discriminant analysis they succeeded in a sensitivity of 74% and a specificity of 96% for the training set. The 33 compounds test set gave an external sensitivity of 70% and a specificity of 92%. The model was built on relatively simple descriptors such as ADME, topology, electronic, energy, surface and geometric descriptors. As the ABCB1 substrate compounds lowered the BBB non-crossing classification rate they therefore were excluded from the final model. Results show that seemingly impermeable compounds, such as cyclosporine A become highly permeable when ABCB1 is saturated.

Table 2: Summary of published classification studies on ABCB1 substrates

Study	N	Algorithm	Descriptor Types	Acc (CV)	N(test)	Ext. Acc.	Ext. SE
Seelig [30]	100		Hydrogen bonding moieties				
Didziapetris et al. [46]	200+1000	C-SAR	ADME desc. (H acc, pKa, nrings, TPSA, MW, etc.)				
Xue et al. [53]	201	SVM	Mol. properties, shape	79.4	13	n.a.	84.2
	201	kNN	Electrotopological state	70.8			
	201	PNN	Quantum chemical prop.	74.4			
	201	C4.5 DT	Geometrical properties	71.5			
Adenot and	91	PLS-DA	ADME, geom.,		33		70.0

Lahana [50]			top., electr., surface, energy descr.				
Gombar et al. [51]	95	SDA	Electrotop., topol., MW, CMR	96.8	58	86.2	94.3
Cabrera et al. [52]	203	LDA	TOPS-MODE desc.	80.5	40	77.5	81.8
De Cerqueira Lima [55]	195	kNN	MolconnZ 4.05 desc	89.0	51	78.0	n.a.
		BQSAR	Atom Pair desc.	90.0	51	72.0	n.a.
		SVM	VolSurf desc.	94.0	51	81.0	n.a.
		BinDT	MOE desc.	86.0	51	66.0	n.a.
Huang et al. [56]	203	SVM	Dragon desc.	80.8	40	90.0	91.0
		ML	(2D + 3D)		40	80.0	82.0
Wang et al. [44]	206	SOM	Atom group, E-state	83.4 80.8	25%	n.a	n.a.
	174	BPNN	Connectivity desc.	84.6 75.3	25%	n.a.	

In 2004 Gombar et al. [51] presented a model based on a set of 95 compounds (63 substrates, 32 non-substrates) assayed with an ABCB1 overexpressing cell line. They used feature selection by stepwise discriminant analysis and developed a two-group linear discriminant function with 27 descriptors. The descriptors include electrotopology, molecular bulk, hydrogen-bond acceptors and donors, lipophilicity, molar refraction and quantified shape attributes of the structure. Their results showed that the size of the molecule, reflected in molecular E-state (MolES), has a significant influence on its properties as substrate and the accumulated data allowed them to postulate the Gombar–Polli Molecular E-state Rule. Molecules with MolES > 110 seem to be substrates and those with MolES < 48 seem to be non-substrates. Their external test set had an overall accuracy of 86.2%, a specificity of 78.3% and a sensitivity of 94.3%. Different descriptors were used by Cabrera and colleagues [52] who used the TOPS–MODE (topological substructural molecular design) approach, which 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) in a region of the molecule. A set of 203 compounds assayed for ABCB1 activity was split in a training set of 163 compounds and a test set. Following the calculation of the descriptors linear discriminant analysis was used and a discriminant function permitting the classification of drugs in substrates and non-substrates

was achieved. The overall accuracy for the test set reached 77.5% with a sensitivity of 81.82% and a specificity of 72.22%. The significance of the bond distance suggests that it could be considered as a more general property providing new information for distinguishing ABCB1 substrates from non-substrates. Acknowledging the complexity of the problem, several authors applied even more sophisticated methods for classification of ABCB1 substrates. Xue and colleagues [53] used support vector machine (SVM), a method which tries to find the best hyperplane able to separate objects in a multidimensional feature space. One hundred and fifty-nine molecular descriptors for shape, electrotopological states, quantum chemical properties and geometrical properties were selected and subsequently recursive feature elimination was applied. With an overall accuracy of 79.4% in the fivefold cross validation set, a sensitivity of 84.2%, and a specificity of 66.7%, the results are slightly better than Penzotti's pharmacophore model [31]. The authors also compared the SVM approach with other methods such as k -nearest neighbor (kNN), probabilistic neural network and C4.5 decision tree but no further improvement was achieved. In another study the authors [54] compared three data sets (ABCB1, human intestinal absorption of molecules and agents causing torsade de pointes) with regard to the use of feature selection before SVM employment. They could show significant improvement in cross-validated accuracy for two of the three data sets studied. Especially for ABCB1 an 11% increase in cross-validated accuracy from 69.3 to 79.4% could be achieved. They also presented their data using the Matthews correlation coefficient, which is a discrete version of Pearson's correlation coefficient and can be used even if the classes are of very different sizes. Wang et al. [44] used both supervised back propagation neural networks and unsupervised Kohonen self-organising maps (SOMs). From the literature a total of 206 chemicals (substrates and inhibitors) were collected and 287 molecular descriptors were calculated. Stepwise discriminant analysis led to 11 relevant descriptors, which resulted in a sensitivity of 83.3% and a specificity of 80.8% using SOMs. De Cerqueira Lima [55] and colleagues used the Penzotti [31] data set (195 diverse substrates and nonsubstrates) to combine different methods including k NN, classification decision tree, Binary QSAR and SVMs and descriptor sets including molecular connectivity indices (MolconnZ), atom pair, VolSurf and molecular operation environment descriptors. Classification studies were carried out separately for each method and descriptor type resulting in a total of 16 combinations. The best models were obtained by SVM classification combined with atom pair descriptors or VolSurf descriptors. All models had higher accuracy in classifying non-substrates than substrates. The most promising approach has been introduced by Huang and colleagues in 2007 [56] with an overall accuracy of 88.6% for the test set. The data set from Cabrera et al. [52] was used and 929 molecular descriptors were calculated. After eliminating redundant information the remaining 79 descriptors were subject to a feature selection using the particle swarm algorithm. Combination with a SVM produced the best predictive model so far using only seven non-correlated and simple descriptors (three constitutional descriptors, two functional group counts and two molecular properties). As suggested by previous studies molecular mass, hydrogen bonds and polar surface area play important roles in substrate recognition of ABCB1 but this study also proposes features such as the number of ring tertiary C atoms and the number of substituted benzene C atoms to be of relevance. In summary, the application of

machine learning methods seems to be the most promising approach for reliable ABCB1 substrate prediction. Among the available machine learning methods the SVM showed the best results so far. Combination with feature selection algorithms are in general more successful and also end up with simple and interpretable descriptors.

Structure-based approaches

High resolution structures of full transporters are difficult to obtain as dynamic membrane proteins are highly problematic to produce, purify and crystallise. ABCB1 was the first multi-drug transporter for which low-to-medium resolution. data were obtained. Publication of the first structure of a full-length ABC-transporter, the lipidA transporter MsbA from Escherichia coli at 4.5 Å in 2001 [57], and the subsequent appearance of the transporter from Vibrio cholerae at 3.8 Å in 2003 [58] seemed to open a path for protein homology modelling of ABCB1. However, the structures had to be withdrawn due to errors in data post processing [59]. Recently, both the corrected structures of MsbA [60] as well as two structures of a putative multi-drug transporter from methicillin resistant Staphylococcus aureus (Sav1866) have been published [61,62] . Especially the structure of Sav1866, which exhibits a higher resolution (3.0 Å) than the new MsbA structure (5.5 Å), has been immediately used as a template structure for the generation of protein homology models. Up to now models for ABCB1 [63] , ABCG2 [64] and ABCC5 [65] have been published. However, Sav1866 has been crystallised in the posthydrolytic state, thus the substrate binding site is facing towards the aqueous chamber. Additionally, the problem of the low sequence similarity in the transmembrane domains remains and protein homology models have to be used very cautiously when using them as starting points for structure-based design attempts.

Conclusion

In conclusion, although significant progress has been made in the past years, these are largely confined to small homogenous data sets. To improve the predictivity of in silico models for ABC-transporter substrates for a broad, diverse data set, new studies using large diverse data sets, such as of Gleeson et al. [49], have to be performed. A major contribution will come from more comparable biological assays between the different laboratories and consistent data interpretation. With our increasing knowledge on protein networks and the complexity of regulatory pathways and protein/protein interaction, new algorithms and descriptors need to be developed especially designed to target polyspecific drug/protein interactions. Last but not least all the systems based information has to be projected back to the molecular level to guide the medicinal chemist in the lead optimisation process to design safer and more efficacious drugs.

Expert Opinion

Pharmaceutical Industry is now facing the problem of decreasing numbers of new chemical entities reaching the market although investments into drug discovery and development reach all time high levels each year. Among others, lack of efficacy and insufficient safety have been identified as main bottlenecks [66]. These are targeted at present under the framework of the Innovative Medicines Initiative, a pan-European 2 billion Euro effort to overcome the main bottlenecks in drug discovery and development. Besides ‘classical’ off-pharmacologies

such as the hERG potassium channel also cytochromes, nuclear receptors and the ABC-transporters are increasingly recognised as being linked to safety sciences. The ABC-transporters are not only involved in drug resistance but are also responsible for the export of a wide variety of drugs from the epithelia of certain tissues (e.g., the intestine, the breast or the blood–brain barrier) into the corresponding lumina. Especially for drugs that are characterised by a low passive absorption rate, this can result in a very limited bioavailability. Furthermore, drug/drug and nutrient/drug interaction at distinct transporters can lead to severe side effects owing to altered plasma levels. Within the past decade the field of ABC-transporters faced a paradigm shift from inhibitor design towards substrate prediction. For a comprehensive overview on the impact of transporters and other antitargets (e.g., hERG, CYP450) on safety and toxicity, the reader is referred to the recently published book on antitargets edited by Vaz and Klabunde [67]. As outlined in this article a variety of computational models has been published, ranging from pharmacophore models to machine learning based models for the classification of ABC-transporter substrates. As the interaction pattern of ABC-transporters and their ligands is polyspecific, both the choice of the proper descriptor set and the classification method becomes an important criterion. Furthermore, several reports proved the value of applying feature selection algorithms in a high dimensional descriptor space. However, owing to the very limited and in part contradictory data available in the literature the problem of the applicability domain of the models published remains to be solved. Current models show good performance in their local chemical space but lack general applicability. This may be overcome by methods and algorithms that take into account the complex and fuzzy type of drug/transporter interaction pattern. We recently introduced the concept of similarity based SAR (SIBAR) and proved its applicability for prediction of ABCB1 inhibitory activity (Figure 3A) [68,69]. Briefly, the calculation of SIBAR descriptors relies on the definition of a reference compound set using the principleof maximum diversity of the compounds. Subsequently, the similarity of the compounds under investigation to these reference compounds is calculated and these similarity values serve as descriptors for QSAR analyses and classification. SIBAR has been shown to provide first positive results in early ADME profiling and also in PLS-QSAR studies of propafenone-type MDR modulators. Recent extensions to shape similarity calculations (3D-SIBAR) gave even better results, however, with the price of higher computational costs. In addition, also descriptors based on topological distribution of atom feature pairs, such as Shannon Entropy Descriptors (SHED) [70,71], have been successfully applied in the field of promiscuous proteins (Figure 3B). The group of Jordi Mestres applied SHED descriptors for a pharmacological profiling of the nuclear hormone receptor family. Nuclear receptors are ligand-dependent transcription factors that in part (e.g., CAR and PXR) also show a promiscuous ligand recognition pattern. These are only two examples showing that there is still the need for developing new descriptors to especially challenge polyspecific proteins such as ABC-transporters.

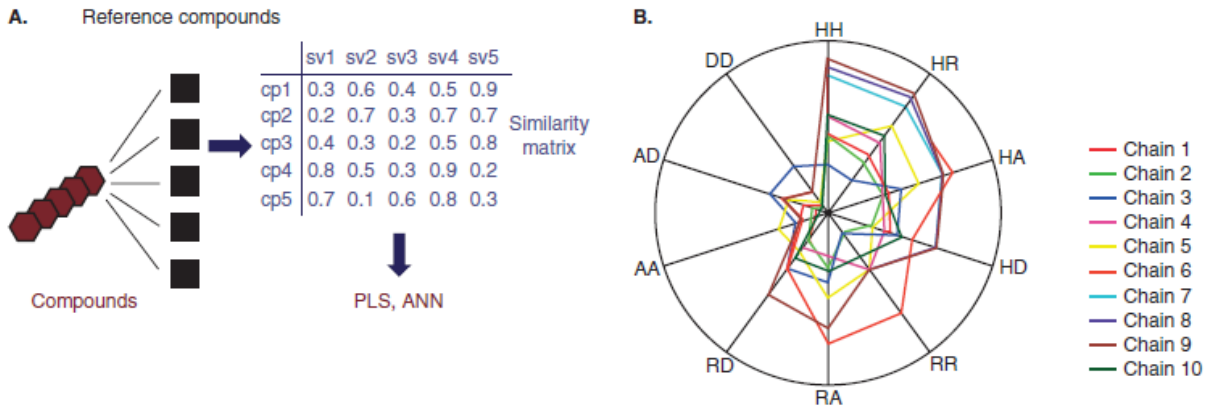


Figure 3: New descriptor types that rely on the chemical similarity principle.

Advanced experimental molecular techniques and in silico pharmacology approaches [72,73] highlight the tight protein interaction network within the ABC-transporter family itself and with other proteins involved in the ADMET process, such as nuclear hormone receptors and cytochrome P450 enzymes. However, systems biology driven drug discovery needs new tools to map the chemical space onto a whole network of proteins. First attempts have recently been presented by Paolini et al. [74] and Keiser et al. [75]. Paolini et al. globally mapped the pharmacological space by constructing a ligand-target matrix that enabled them to explore the relationships between ligand structure and biological target proteins. Furthermore, they also provide a comprehensive listing of molecular properties of highly related and distinct gene family ligands and consecutively observe that the molecular properties of ligands are correlated with their target class. Keiser and colleagues use a slightly different approach in relating proteins based on the chemical similarity of their ligands, which not only enables them to assess the 'druggability' of certain proteins but also allows the prediction of biological fingerprints of small molecules. Applying these approaches to the network of ABC-transporters might be a very interesting and promising exercise. It is important to mention that X-ray crystallographic structures of bacterial homologues of the ABC-transporter family are available. These are the structure of SAV1866 from S. aureus [61] and the updated and refined MsbA structures from E. coli, V. cholera, and Salmonella typhimurium [60]. Protein homology modelling studies can support and complement the ligand-based studies in terms of predicting binding site topology and binding mode of substrates and inhibitors of ABC-transporters. However, most of the published crystal structures harbour a nucleotide as ligand and therefore resemble the posthydrolytic state of the transporter. The recently corrected MsbA structure from V. cholerae seems to represent the only X-ray crystal structure representing a non-energised conformation and therefore seems to be the most suitable template for homology modelling and substrate binding site identification. Unfortunately, this crystal structure shows a rather low resolution of 5.5 Å and contains only the C α -atoms of the protein backbone. The community definitely needs and desperately awaits better structures at higher resolution to pursue structure-based design approaches.

Finally, it has to be noted that at present only very limited data are available which guide the medicinal chemist in the lead optimisation process. Despite the very early studies of Klopman [76], who analysed several hundred inhibitors of ABCB1 and derived a set of pharmacophoric

and pharmacophobic sub structures almost no information can be found in the literature that exemplifies successful attempts to decrease ABC-transporter substrate properties in a lead compound series. Thus, according to Raub [77] the best approach to diminish the limiting effects of ABCB1 on a particular scaffold for now is to increase passive diffusion. Then the pump can no longer maintain a concentration gradient owing to high and rapid passive diffusion. However, recently scientists from Merck Research Laboratories published two studies in which they designed in and designed out ABCB1 substrate properties. In a series of kinesin spindle protein inhibitors efflux by ABCB1 was reduced through modulation of the basicity of the nitrogen atom by β -fluorination [78]. On the other hand, incorporation of an extra hydrogen-bond donor rendered a series of antagonists of the Kv1.5 potassium channel ABCB1 substrates, thus reducing brain exposure [79].

Acknowledgement

This work was supported by grants from the Austrian Promotion Agency (B1-812074) and from the Austrian Science Fund (L344-N17).

References

Papers of special note have been highlighted as either of interest (•) or of considerable interest (••) to readers.

1. Mueller M. Available from:http://nutrigene.4t.com/humanabc.htm

2. Gottesman MM, Fojo T, Bates SE. Multidrug resistance in cancer: role of ATP-dependent transporters. Nat Rev Drug Discov 2002 ; 2 : 48 -58

3. Ross DD, Doyle LA. Mining our ABCs: pharmacogenomic approach for evaluating transporter function in cancer drug resistance. Cancer Cell 2004 ; 6 : 105 -7

4. Schinkel AH, Jonker JW. Mammalian drug efflux transporters of the ATP binding cassette (ABC) family: an overview. Adv Drug Deliv Rev 2003 ; 55 : 3 -29

5. van Tellingen O. The importance of drug-transporting P-glycoproteins in toxicology. Toxicol Lett 2001 ; 120 : 31 -41

6. Thiebaut F, Tsuruo T, Hamada H, et al. Cellular localization of the multidrug-resistance gene product P-glycoprotein in normal human tissues. Proc Natl Acad Sci USA 1987 ; 84 : 7735 -8

7. Schinkel AH, Smit JJM, van Tellingen O, et al. Disruption of the mouse mdr1a P-glycoprotein gene leads to a defi ciency in the blood-brain barrier and to increased sensitivity to drugs. Cell 1994 ; 74 : 491 -502

8. Pal D, Mitra AK. MDR- and CYP3A4-mediated drug–drug interactions. J Neuroimmune Pharmacol 2006 ; 1 : 323 -39

9. Kiani J, Imam SZ. Medicinal importance of grapefruit juice and its interaction with various drugs. Nutr J 2007 : 6 -33

10. Wu C-Y, Benet LZ. Predicting drug disposition via application of BCS: transport/absorption/elimination interplay and development of a biopharmaceutics drug disposition classification system. Pharm Res 2005 ; 22 : 11 -33

11. Pauli-Magnus C, Meier PJ. Hepatobiliary transporters and drug-induced cholestasis. Hepatology 2006 ; 44 : 778 -87

12. Szakacs G, Patterson J, et al. Targeting multidrug resistance in cancer. Nat Rev Drug Discov 2006 ; 5 : 219 -33

13. Feng B, Mills J, Davidson R, et al. In vitro P-glycoprotein assays to predict the in vivo interactions of P-glycoprotein with drugs in the central nervous system. Drug Metab Dispos 2008 ; 36 (2): 268 -75

14. Garrigos M, Belehradek J, Mir LM, Orlowski S. Absence of cooperativity for MgATP and verapamil effects on the ATPase activity of P-glycoprotein containing membrane vesicles. Biochem Biophys Res Commun 1993 ; 196 : 1034 -41

15. Polli JW, Wring SA, Humphreys JE, et al. Rational use of in vitro P-glycoprotein assays in drug discovery. J Pharmacol Exp Ther 2001 ; 299 : 620 -8

16. Schmid D, Ecker G, Kopp S, et al. Structure-activity relationship studies of propafenone analogs based on P-glycoprotein ATPase activity measurements. Biochem Pharmacol 1999 ; 58 : 1447 -56

17. Hochman JH, Yamazaki M, Ohe T, JH L. Evaluation of drug interactions with P-glycoprotein in drug discovery: in vitro assessment of the potential for drug-drug interactions with P-glycoprotein. Curr Drug Metab 2002 ; 3 : 257 -73

18. Tiberghien F, Loor F. Ranking of P-glycoprotein substrates and inhibitors by a calcein-AM fl uorometry screening assay. Anticancer Drugs 1996 ; 7 : 568 -78

19. Zhou H, Wu S, Zhai S, et al.Design, synthesis, cytoselective toxicity, structure-activity relationships, and pharmacophore of thiazolidinone derivatives targeting drug-resistant lung cancer cells. J Med Chem 2008 ; 51 : 1242 -51

20. Szakacs G, Annereau J-P, Lababidi S, et al. Predicting drug sensitivity and resistance: profiling ABC transporter genes in cancer cells. Cancer Cell 2004 ; 6 : 129 -37

21. Safa AR, Roberts S, Agresti M, Fine RL. Tamoxifen aziridine, a novel affi nity probe for P-glycoprotein in multidrug-resistant cells. Biochem Biophys Res Commun 1994 ; 202 : 606 -12

22. Schuetz EG, Beck WT, Schuetz JD. Modulators and substrates of P-glycoprotein and cytochrome P4503A coordinately up-regulate these proteins in human colon carcinoma cells. Mol Pharmacol 1996 ; 49 : 311 -8

23. Zhang X, Ritke M, Yalowich J, et al. P-glycoprotein mediates profound resistance to bisantrene. Oncol Res 1994 ; 6 : 291 -301

24. Humblet C, Marshall GR. Pharmacophore modelling and receptor mapping. Ann Rep Med Chem 1980 ; 15 : 267 -76

25. Wolber G, Seidel T, Bendix F, Langer T. Molecule-pharmacophore superpositioning and pattern matching in computational drug design. Drug Discov Today 2008 ; 13 : 23 -9

26. Chang C, Ekins S, Bahadduri P, Swaan PW. Pharmacophore-based discovery of ligands for drug transporters. Adv Drug Deliv Rev 2006 ; 58 : 1431 -50

• Review summarising the current success stories on pharmacophore modelling for drug transporters. It covers ABC-transporters, neurotransmitter transporters and human proton-dependent intestine transporters.

27. Pearce HL, Winter MA, Beck WT. Structural characteristics of compounds that modulate P-glycoprotein-associatedn multidrug resistance. Adv Enzyme Regul 1989 ; 30 : 357 -73

28. Pearce HL, Safa AR, Bach NJ, et al. Essential features of the P-glycoprotein pharmacophore as defi ned by a series of reserpine analogs that modulate multidrug resistance. Proc Natl Acad Sci USA 1989 ; 86 : 5128 -32

29. Ecker G, Huber M, Schmid D, Chiba P. The importance of a nitrogen atom in modulators of multidrug resistance. Mol Pharmacol 1999 ; 56 : 791 -6

30. Seelig A. A general pattern for substrate recognition by P-glycoprotein. Eur J Biochem 1998 ; 251 : 252 -61

31. Penzotti JE, Lamb ML, Evensen E, Grootenhuis PDJ. A computational ensemble pharmacophore model for identifying substrates of P-glycoprotein. J Med Chem 2002 ; 45 : 1737 -40

32. Ekins S, Kim RB, Leake BF, et al. Application of three-dimensional quantitative structure-activity relationshipsof P-glycoprotein inhibitors and substrates. Mol Pharmacol 2002 ; 61 : 974 -81

33. Ekins S, Bravi G, Binkley S, et al. Three- and four-dimensional quantitative structure activity relationship analyses of cytochrome P-450 3A4 inhibitors. J Pharmacol Exp Ther 1999 ; 290 : 429 -38

34. Ekins S, Bravi G, Wikel JH, Wrighton SA. Three-dimensional-quantitative structure activity relationship analysis of cytochrome P-450 3A4 substrates. J Pharmacol Exp Ther 1999 ; 291 : 424 -33

35. Pajeva IK, Wiese M. Pharmacophore model of drugs involved in P-glycoprotein multidrug resistance: explanation of structural variety (Hypothesis). J Med Chem 2002 ; 45 : 5671 -86

36. Ekins S, Kim RB, Leake BF, et al. Three-dimensional quantitative structure-activity relationships of inhibitors of P-glycoprotein. Mol Pharmacol 2002 ; 61 : 964 -73

37. Garrigues A, Loiseau N, Delaforge M, et al. Characterization of two pharmacophores on the multidrug transporter P-glycoprotein. Mol Pharmacol 2002 ; 62 : 1288 -98

38. Cianchetta G, Singleton RW, Zhang M, et al. A pharmacophore hypothesis for P-glycoprotein substrate recognition using GRIND-based 3D-QSAR. J Med Chem 2005 ; 48 : 2927 -35

39. Langer T, Eder M, Hoffmann RD, et al. Lead identification for modulators of multidrug resistance based on in silico screening with a pharmacophoric feature model. Arch Pharm (Weinheim) 2004 ; 337 (6): 317 -27

40. Chang C, Bahadduri PM, Polli JE, et al. Rapid identifi cation of P-glycoprotein substrates and inhibitors. Drug Metab Dispos 2006 ; 34 : 1976 -84

41. Chang C, Ekins S, Bahadduri P, Swaan PW. Pharmacophore-based discovery of ligands

for drug transporters. Adv Drug Deliv Rev 2006 ; 58 (12-13): 1431 -50

42. Li WX, Li L, Eksterowicz J, et al. Significance analysis and multiple pharmacophore models for differentiating P-glycoprotein substrates. J Chem Inf Model 2007 ; 47 : 2429 -38

•• Application of ensemble pharmacophore modelling in conjunction with a classification tree.

43. Cramer J, Kopp S, Bates SE, et al. Multispecificity of drug transporters: probing inhibitor selectivity for the human drug effl ux transporters ABCB1 and ABCG2. ChemMedChem 2007 ; 2 : 1783 -8

•• Describes selectivity profi ling of ABCB1 and ABCG2 inhibitors of the propafenone-type by quantitative and qualitative pharmacophore modelling.

44. Wang YH, Li Y, Yang SL, Yang L. Classification of substrates and inhibitors of P-glycoprotein using unsupervised machine learning approach. J Chem Inf Model 2005 ; 45 : 750 -7

45. Seelig A, Landwojtowicz E. Structure-activity relationship of P-glycoprotein substrates and modifi ers. Eur J Pharm Sci 2000 ; 12 : 31 -40

46. Didziapetris R, Japertas P, Avdeef A, Petrauskas A. Classification analysis of P-glycoprotein substrate specifi city. J Drug Target 2003 ; 391 -406

• Classification study that defines the 'rule of four' for ABCB1 substrates by C-SAR.

47. Available from: http://pharma-algorithms.com/adme_boxes.htm

48. Didziapetris R, Japertas P, Riauba L, et al.Classification SAR (C-SAR) in predictionof P-glycoprotein substrate specifi city. Poster Abstr EuroQSAR 2002

49. Gleeson MP. Generation of a set of simple, interpretable ADMET rules of thumb. J Med Chem 2008 ; 51 (4): 817 -34

•• Derives simple ADMET rules for > 30,000 GSK in-house compounds for 15 assays (among those: ABCB1 efflux, hERG liability, inhibition of various cytochrome isoenzymes, CNS penetration and solubility) based on only four descriptors.

50. Adenot M, Lahana R. Blood-brain barrier permeation models: discriminating between potential CNS and non-CNS drugs including P-glycoprotein substrates. J Chem Inf Comput Sci 2004 ; 44 : 239 -48

51. Gombar VK, Polli JW, Humphreys JE, et al. Predicting P-glycoprotein substrates by a quantitative structure-activity relationship model. J Pharm Sci 2004 ; 93 : 957 -68

• Classification study that defines the 'Gombar–Polli-rule' for ABCB1 substrates based on molecular E-state values.

52. Cabrera MA, González I, Fernàndez C, et al. A topological substructural approach for the prediction of P-glycoprotein substrates. J Pharm Sci 2006 ; 95 : 589 -606

53. Xue Y, Yap CW, Sun LZ, et al. Prediction of P-glycoprotein substrates by a support vector machine approach. J Chem Inf Comput Sci 2004 ; 44 : 1497 -505

54. Xue Y, Li ZR, et al. Effect of molecular descriptor feature selection in support vector machine classifi cation of pharmacokinetic and toxicological properties of chemical agents. J Chem Inf Model 2004 ; 44 : 1630 -8

55. De Cerqueira Lima P, Golbraikh A, Oloff S, et al. A: combinatorial QSAR modeling of P-glycoprotein substrates. J Chem Inf Model 2006 ; 46 : 1245 -54

56. Huang J, Ma G, Muhammad I, Cheng Y. Identifying P-glycoprotein substrates using a support vector machine optimized by a particle swarm. J Chem Inf Model 2007 ; 47 : 1638 -47

•• Classification of ABCB1 substrates and non-substrates by using a particle swarm algorithm as wrapper for a SVM classifi er.

57. Chang G, Roth CB. Structure of MsbA from E. coli: a homolog of the multidrug resistance ATP binding cassette (ATP) transporters. Science 2001 ; 293 : 1793 -800

58. Chang G. Structure of MsbA form Vibrio Cholerae: a multidrug resistance ABC transporter in a closed conformation. J Mol Biol 2003 ; 330 : 419 -30

59. Chang G. Retraction. Science 2006 ; 314 : 1875b

60. Ward A, Reyes CL, Yu J, et al. Flexibility in the ABC transporter MsbA: alternating access with a twist. Proc Natl Acad Sci USA 2007 ; 104 : 19005 -10

61. Dawson RJP, Locher KP. Structure of a bacterial multidrug ABC transporter. 2006 ; 443 : 180 -5

•• Presents the first X-ray structure of a bacterial ABC-transporter that is a homologue to human ABC-transporters in the outward-facing conformation inthe presence of ATP.

62. Dawson RJP, Locher KP. Structure of the multidrug ABC transporter Sav1866 from Staphylococcus aureus in complex with AMP-PNP. FEBS Lett 2007 ; 581 : 935 -8

63. Ravna AW, Sylte I, Sager G. Molecular model of the outward facing state of the human P-glycoprotein (ABCB1), and comparison to a model of the human MRP5 (ABCC5). TBiomed 2007 ; 4 : 33

64. Hazai E, Bikádi Z. Homology modelling of breast cancer resistance protein (ABCG2). J Struct Biol 2008 ; 162 : 63 -74

65. Ravna AW, Sylte I, Sager G. A molecular model of a putative substrate releasing conformation of multidrug resistance protein 5 (MRP5). Eur J Med Chem In Press, Corrected Proof

66. The Innovative Medicines Initiative. Available from: http://imi.europa.eu

67. Vaz R, Klabunde T. Antitargets: prediction and prevention of drug side effects.Edition 1. Weinheim: Wiley-VCH; 2008

68. Zdrazil B, Kaiser D, Kopp S, et al. Similarity-Based Descriptors (SIBAR) as tool for QSAR studies on P-glycoprotein inhibitors: infl uence of the reference set. QSAR Comb Sci 2007 ; 26 : 669 -78

69. Klein C, Kaiser D, Kopp S, et al. Similarity based SAR (SIBAR) as tool for earlyADME profiling. J Comput Aided Mol Des 2002 ; 16 : 785 -93

70. Gregori-Puigjane E, Mestres J. SHED: Shannon Entropy Descriptors from topological feature distributions. J Chem Inf Model 2006 ; 46 : 1615 -22

71. Mestres J, Martin-Couce L, Gregori-Puigjane E, et al. Ligand-based approach to in silico pharmacology: nuclearreceptor profiling. J Chem Inf Model 2006 ; 46 : 2725 -36

72. Ekins S, Mestres J, Testa B. In silico pharmacology for drug discovery: methods for virtual ligand screening and profi ling. Br J Pharmacol 2007 ; 152 : 9 -20

• Review summarising the current in silico methods for ligand screening and profiling.

73. Ekins S, Mestres J, Testa B. In silico pharmacology for drug discovery: applications to targets and beyond. Br J Pharmacol 2007 ; 152 : 21 -37

• Review illustrating the current state-of-the-art methods used in target-based drug design.

74. Paolini GV, Shapland R, van Hoorn W, et al. Global mapping of pharmacological space. Nat Biotechnol 2006 ; 24 : 243 -57

75. Keiser MJ, Roth BL, Armbruster BN, et al. Relating protein pharmacology by ligand chemistry. Nat Biotech 2007 ; 25 : 197 -206

•• Applies a chemo-centric approach to relate ligands to their molecular targets based on ligand similarity.

76. Klopman G, Shi LM, Ramu A. Quantitative structure-activity relationship of multidrug resistance reversal agents. Mol Pharmacol 1997 ; 52 : 323 -34

77. Raub TJ. P-glycoprotein recognition of substrates and circumvention through rational drug design. Mol Pharm 2006 ; 3 : 3 -25

78. Cox CD, Breslin MJ, Whitman DB, et al. Kinesin spindle protein (KSP) inhibitors. Part V: discovery of 2-propylamino-2,4-diaryl-2,5-dihydropyrroles as potent, water-soluble KSP inhibitors, and modulation of their basicity by [beta]-fl uorination to overcome cellular efflux by P-glycoprotein. Bioorg Med Chem Lett 2007 ; 17 : 2697 -702

•• Describes the experimental rendering on a rational basis of substrate/non-substrate properties for a series of compounds targeting kinesin spindle proteins.

79. Nanda KK, Brad Nolt M, Cato MJ, et al.Potent antagonists of the Kv1.5 potassium channel: synthesis and evaluation of analogous N,N-diisopropyl-2-(pyridine-3-yl)acetamides. Bioorg Med Chem Lett 2006 ; 16 : 5897 -901

Author Contribution

This chapter was published in *Expert Opinion on Drug Metabolism and Toxicology* as

Demel M., Schwaha R., Krämer O., Ettmayer P., Haaksma E., Ecker G.F. **In Silico Prediction of Substrate Properties for ABC-Multidrug-Transporters.** *Expert Opinion on Drug Metabolism and Toxicology, 2008, 4(9):1167-1180.*

The initial manuscript for publication was written by me and R. Schwaha. Mag. Schwaha wrote the part reviewing the classification studies. The section “Structure-based approaches” was written by me and Mag. Schwaha. Dr. Krämer, Dr. Ettmayer, Dr. Haaksma and especially Prof. Ecker contributed substantially to the “Expert Opinion” section. All the other sections were written by me. The manuscript was reviewed, edited and improved by all the co-authors. Prof. Ecker supervised this project.

Addendum: More Ligand-based Efforts to Characterize ABC-transporter ligands

The preceding chapter was published in 2008. At least, three more ligand-based studies that have been published since then also require proper appraisal. First of all, Broccatelli et al. investigated the influence of ABC-transporter-mediated efflux on CNS side effects [1]. They examined 64 marketed drugs (H_1-blockers, antidepressants and antipsychotics) in order to explore the relationship between sedation (a H_1-receptor mediated side effect) and orthostatic hypotension (a α_1-mediated side effect). They concluded that these two clinical relevant side effects are not results of the passive diffusion of these molecules, but are highly related to ABCB1-mediated efflux at the blood-brain-barrier. Compounds such as mizolastine and aelastine are not effluxed by ABCB1 and contribute therefore to sedation and orthostatic hypotension when given and high doses. This study highlights that ABCB1 must be considered indeed as a drug target for compounds with their primary mode of action outside the CNS.

In another study Wang and colleagues present a SVM model for ABCB1 substrates on basis of a comprehensive data set consisting of 332 distinct structures [2]. The chemical structures were encoded by autocorrelation descriptors, descriptors contained in the Molecular operating Environment software package and by EFCP_4 fingerprints. The analysis of their model points out the importance of van-der-Waals surface area (VSA) descriptors. Furthermore, they could associate the nitrile and the sulfoxide groups with a higher frequency in non-substrates.

At last the work of LeDonne et al. requires to be acknowledged [3]. They used proprietary compounds from AstraZeneca assayed in a MDCK-MDR1 *in-vitro* model. Furthermore, they make use of the SALI measure proposed by Guha and Van Drie to assess the reliability of their predictive models [4], [5]. Briefly, SALI is an estimate designed to improve the understanding of the applicability domain of predictive models and is based on the concept of activity cliffs. The authors found that SALI can substantially improve the evaluation of model performance and they recommend to use this approach in the evaluation of prediction models.

Bibliography

[1] F. Broccatelli, E. Carosati, G. Cruciani, and T. I. Oprea, “Transporter-mediated Efflux Influences CNS Side Effects: ABCB1, from Antitarget to Target.,” *Molecular Informatics*, vol. 29, no. 1–2, pp. 16–26, Jan. 2010.

[2] Z. Wang, Y. Chen, H. Liang, A. Bender, R. C. Glen, and A. Yan, “P-glycoprotein substrate models using support vector machines based on a comprehensive data set.,” *Journal of Chemical Information and Modeling*, vol. 51, no. 6, pp. 1447–56, Jul. 2011.

[3] N. C. LeDonne, K. Rissolo, J. Bulgarelli, and L. Tini, “Use of structure-activity landscape index curves and curve integrals to evaluate the performance of multiple machine learning prediction models.,” *Journal of Cheminformatics*, vol. 3, no. 1, p. 7, Jan. 2011.

[4] R. Guha and J. H. Van Drie, “Structure-activity landscape index: identifying and quantifying activity cliffs.,” *Journal of Chemical Information and Modeling*, vol. 48, no. 3, pp. 646–58, Mar. 2008.

[5] R. Guha and J. H. Van Drie, “Assessing how well a modeling protocol captures a structure-activity landscape.,” *Journal of Chemical Information and Modeling*, vol. 48, no. 8, pp. 1716–28, Aug. 2008.

Chapter 10: Publication II - Predicting Ligand Interactions with ABC Transporters in ADME

Abstract

ABC-type drug efflux pumps, e.g., ABCB1 (=P-glycoprotein, 0MDR1), ABCC1 (=MRP1), andABCG2 (=MXR, =BCRP), confer a multi-drug resistance (MDR) phenotype to cancer cells. Furthermore, the important contribution of ABC transporters for bioavailability, distribution, elimination, and blood–brain barrier permeation of drug candidates is increasingly recognized. This review presents an overview on the different computational methods and models pursued to predict ABC transporter substrate properties of drug-like compounds. They encompass ligand-based approaches ranging from "simple rule"-based efforts to sophisticated machine learning methods. Many of these models show excellent performance for the data sets used. However, due to the complex nature of the applied methods, useful interpretation of the models that can be directly translated into chemical structures by the medicinal chemist is rather difficult. Additionally, very recent and promising attempts in the field of structure-based modeling of ABC transporters, which embody homology modeling as well as recently published X-ray structures of murine ABCB1, will be discussed.

Introduction: ABC Transporters and ADMET profiling.

Within the past years, increasing attention has been drawn to include membrane transporters, in addition to classical drug metabolism studies, in the early ADMET (absorption, distribution, metabolism, excretion, and toxicity) investigation phase. Although a variety of transmembrane transporter proteins have been shown to influence ADMET [1– 3], a considerable amount of attention is focusing on multidrug exporters of the multidrugresistant ATP-binding cassette (ABC) transporter family [4]. Many of the cytostatic drugs that are currently in clinical use in cancer therapy have been shown to be substrates of these transporters. Transporter overexpression results in multidrug resistance (MDR) and consequently in reduced chemotherapeutic efficacy [5]. Although the role of ABC transporters in cancer is still a matter of intense research (especially in the area of cancer stem cells; for details, see Dean et al. [6]), it is nowadays widely accepted that ABC proteins also participate in chemo-defense mechanisms at cellular or tissue barriers and, therefore, are also responsible for both limited bioavailability as well as drug – drug interactions [7]. The human ABC transporter ABCB1 (also known as P-glycoprotein (P-gp) and MDR1), which was the first ABC transporter discovered more than 30 years ago [8], represents the paradigm transporter in this field. However, several additional members of the ABC superfamily (ABCC1– 9 as well as ABCG2) have been shown to influence ADMET behavior of drugs. ABCB1 is a 170-kDa protein constituted of two transmembrane domains and two nucleotide-binding domains, which is physiologically expressed mainly in the liver, the kidney, the intestine, the blood– brain barrier, and the blood –placenta barrier. The early identification of ligands interacting with ABCB1 is one of the challenging topics in transporter-mediated ADMET profiling. A striking feature of the aforementioned ABC proteins, and especially of ABCB1, is their polyspecificity (promiscuity) in substrate and inhibitor recognition. In general, there are many reasons for the occurrence of polyspecific interactions among ligands

and their targets [9], such as:

- conformational flexibility of the protein,
- multiple interaction sites at the protein,
- size and complexity of the ligands, and
- "imperfect complementarity" (partial recognition) in the interaction.

From this collection, it can be seen that both interaction partners, the ligands as well as the proteins, contribute to the phenomenon of polyspecificity. Unfortunately, most of the traditional computational drug-design methods can deal with this complexity only at a limited level. Therefore, the current strategies to model the interactions of ABC transporters with their ligands shift from classical QSAR-based approaches (which require homologous series of compounds that interact in a unique binding mode) to more complex methods, ranging from the application of support vector machines (SVMs) and ensemble learning algorithms to multiple pharmacophore modeling. The use of these complex methods, which are in part difficult to interpret, immediately raises the question if such approaches might be accepted in the same way by the medicinal-chemistry community as simple concepts like, e.g., the rule of four postulated by Didziapetris et al. [10]. Here, we present an overview on the current in silico efforts that have been undertaken to elucidate the substrate properties of ABCB1.

Ligand-based in silico Models for ABC Transporter Substrates

Pharmacophore Models

The early in silico efforts to illuminate key interactions of ABC transporters with their ligands were based on pharmacophore modeling [11] [12]. Up to now, many additional studies presented pharmacophore models for ABCB1 substrates [13 – 24]. A comprehensive overview of these approaches is given in Chang et al. [21] and Demel et al. [24]. Overall, the published pharmacophore models show some convergence concerning the features responsible for drug – transporter interactions (hydrophobic and H-bond acceptor features). However, most of them are not directly comparable, due to remarkable differences in the distance matrices of the indicated features. This reflects the overlapping substrate specificity of ABC transporters and suggests that multiple or ensemble pharmacophore models might be a better way to cover the pharmacophore space of ABCB1 substrates. This approach has been undertaken by Li et al. [22], who provided an ABCB1 substrate-classification study based on multiple pharmacophores in conjunction with a decision tree algorithm. This approach resulted in an excellent overall prediction accuracy of 87.7% for a training set of 163 compounds, and 87.6% for a test set of 97 compounds. This combination of multiple pharmacophore models with a classification tree algorithm seems to represent a successful state-of-the-art method to determine substrate properties for ABCB1. To probe the selectivity of two different ABC transporters, ABCB1 and ABCG2, with respect to a series of propafenone-type inhibitors, we could recently demonstrate that the two pharmacophores derived for the two transporters differed mainly in the region of the N-atom.Whereas, in the case of ABCB1, an H-bond acceptor feature was present in this region, for ABCG2 inhibitors this part of the molecule seemed not to be relevant [25]. Interestingly, the ABCG2 pharmacophore seemed to be more tolerant towards chemical modifications, which could also

be demonstrated in Hansch analyses. These first results demonstrate the feasibility of designing in/designing out anti-target interactions. Furthermore, they also demonstrate that, even in case of two polyspecific transporters with partially overlapping substrate profiles, selectivity profiling is possible.

Machine Learning Models

Besides pharmacophore modeling, a variety of machine learning algorithms such as linear discriminant analysis [26], support vector machines (SVMs) [27] [28], and artificial neural networks [29] have been applied to correctly discriminate ABCB1 ligands. These recent studies are reviewed in more detail in Li et al. [30] and Demel et al. [24]. Themost promising approach has been introduced by Huang et al. [28], in 2007, with an overall accuracy of 88.6% for a test set. They used a data set of 203 molecules and calculated 929 molecular descriptors. By combination of an SVM classification with a particle swarm-optimization algorithmas feature selection pre-processing step, a model with excellent predictivity was produced. This model is based on seven non-correlated and simple descriptors (three constitutional descriptors, two functional group counts, and two molecular properties). In convergence with previous studies, their results show that the molecular weight, H-bonds, and the polar surface area play important roles in substrate recognition of ABCB1, as well as features like the number of ring tertiary C-atoms and the number of substituted benzene C-atoms.

"Simple Rule" Models for the Classification of ABCB1 Substrates.

The abovementioned ligand-based models provide in general an excellent classification performance, but suffer in most cases from lack of interpretability. This is mainly due to the highly advanced nature of the algorithms (e.g., SVMs and neural networks) or the huge amount of complex descriptor types (e.g., Molconn-Z, VolSurf, and shape-based descriptors) used in these studies. Besides these highly complex models, also attempts were made to achieve simple and fast interpretation of the molecular features responsible for substrate properties for ABCB1. These studies mainly focus on deriving "simple rules of thumb" that can easily guide the medicinal chemist in a hit-to-lead optimization program [10] [26] [31] [32]. The first attempt to classify ABCB1 substrates and non-substrates was made by Seelig, in 1998, on a set of 100 structurally diverse 3D molecules. Setting the focus on the importance of N- and O-atoms (=H-bond acceptors, =electron-donor groups), Seelig concluded that substrate recognition is particularly based on one of two specific H-bonding patterns. Her analysis suggests that substrates contain either two H-bonding features in a spatial separation of ca. 2.5 _ or three H-bonding features with a spatial separation of the outer two features of ca. 4.6 _. She further showed a correlation between H-bonding strength and ABCB1 binding strength and defined the partitioning into the membrane as the rate-limiting step [31]. Didziapetris et al. postulated the _rule of four_, which states that compounds with a number of H-bond acceptors (NþO)_8, a molecular weight (MW)>400 Da, and with a pKa>4 for the most acidic group are likely to be ABCB1 substrates, whereas compounds with (NþO)_4, MW<400, and with a pKa<8 for the most basic group have a higher probability to be non-substrates [10]. In 2004, Gombar et al. postulated the "Gombar–Polli Molecular E-state (MolES) Rule" based on a set of 95 compounds (63 substrates, 32

non-substrates) assayed in an ABCB1 overexpressing cell line [26]. They developed a two-group linear discriminant function with 27 descriptors. Molecules with MolES>110 have a high probability for being substrates and those with MolES<48 seem to be non-substrates. Their external test set consisting of 58 compounds had an overall accuracy of 86.2%, a specificity of 78.3%, and a sensitivity of 94.3%. In a recent publication, Gleeson used 1975 compounds with measured ABCB1 efflux ratios of a proprietary GlaxoSmithKline (GSK) database [32]. With only four simple ADMET descriptors (molecular weight, log P value, positive ionization state, and negative ionization state) it was demonstrated that neutral or basic molecules showing aMW>400 and a log P value>4 are more likely to be transported by ABCB1 than acidic or zwitterionic compounds. The results of these four "simple rules" developed so far are summarized in Table 1.

Table 1: Summary of Published „Simple Rules" for ABCB1 Substrates

Reference	No. compounds	Availability of structures	Type	Substrate if ("rule")	External validation
Seelig [31]	100	Yes	Analysis of 3D structures	Spatial separation of H-bond acceptors either 2.5 or 4.6 A	No
Didziapetris et al. [10]	200/1000	No	Stepwise C-SAR	nHBA>8 &&MW>400 Da && pKa>4	No
Gombar et al. [26]	95	Yes	Interpretation of an LDA model based on E-state descriptors	MolES>110	Yes
Gleeson[32]	1975	No	ANOVA	MW>400 && log P> 4 && molecule is neutral or basic	No

Although the four studies mentioned above present "simple rules" that might have a greater impact than more complex (and more predictive) machine learning models, they suffer from the potential drawback that most of these models are either based on a small collection of compounds or that the compounds are not disclosed to the public. In order to overcome these disadvantages, we have generated a large collection of 497 publicly available compounds on the basis of literature compounds as well as on compounds derived from the NCI60 screen on ABC transporters. After encoding our compounds with eleven simple physicochemical descriptors, this data set was subjected to RuleFit modeling as described by Friedman and Popescu [33]. This modelling procedure allows the generation of an ensemble of conjunctive rules (i.e., two or more conditions are connected via logical AND_(&&) operators) in form of a linear model. After careful evaluation of our model, we received a sensitivity of 73% and a specificity of 72%. The most important descriptors in this model were the molar refractivity (mr), the number of H-bond acceptors (a_acc), and the total hydrophobic Van-der-Waals surface area (PEOE_VSA_HYD). The univariate class discriminating power of these three descriptors is confirmed by binned histograms (Fig. 1). The careful interpretation of the conjunctive rules generated by this model suggests that substrates are in general characterized by a mr >10, a high hydrophobicity (PEOE_VSA_HYD > 300 and 5 < log P < 7), and that they are associated with more H-bond acceptor features (which is expressed as a_acc > 7 and vsa_acc > 28) than nonsubstrates. A detailed definition of the derived rules can be found in Table 2.

Table 2: „Rules" for ABCB1 Substrates Derived from RuleFit Modeling

Class	Rule
NS	a_acc≤8 && vsa_acc≤33 && vsa_don≥21
S	vsa_acc≥28.51 && log P≥4.6
NS	mr≤10 && vsa_acc≤20.5 && vsa_don≥8.5
S	PEOE_VSA_HYD≥297.42 && a_acc≥7 && log P≤7
S	mr≥9.11 && a_don≤5

A comparison of the rules provided by our model with previously described rules (listed in Table 1) shows general convergence. Briefly, substrates can be defined as being larger in size, more hydrophobic, and having more H-bond acceptor atoms than non-substrates. Very recently, we additionally attempted to derive a global descriptor set for the three different ADMET-relevant ABC transporters ABCB1, ABCC1, and ABCG2 [34]. The descriptors were selected on the basis of different feature-selection algorithms out of an initial compilation of more than 170 2D descriptors. Although the classification models for all three proteins showed satisfying performance, it was not able to identify an entirely unique feature set that can be globally used to characterize substrates of these transporters. The only

significant feature identified was that nonsubstrates contain a higher number of rigid bonds. This suggests that a decrease in the number of rotatable bonds, or alternatively an increase in the number of rigid bonds, might be a potential strategy to avoid general ABC transporter-mediated bioavailability issues.

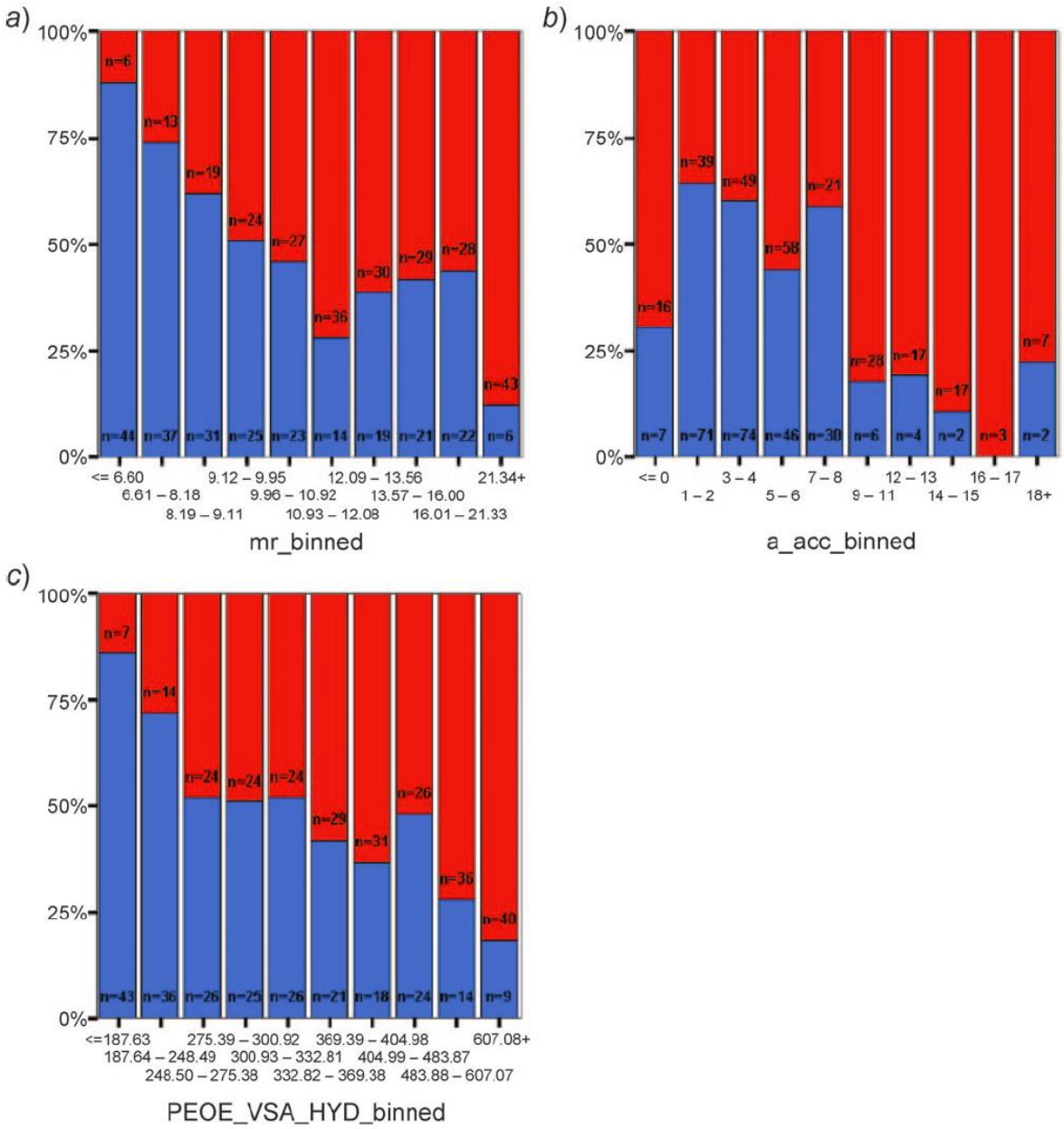


Figure 1: Histogram of the most important descriptors

Structure-Based Models

Human ABCB1 was the first ABC transporter for which low-to-medium resolution data were obtained from cryo-electron microscopy (EM) studies [35 – 37]. These studies provided pioneering insights into the domain architecture of ABCB1, which is constituted of two nucleotide-binding domains (NBDs) and two transmembrane domains (TMDs), and furthermore into changes in domain positions upon different stages of the transport cycle. The publication in 2001 of the first bacterial homologue structure of a full length ABC transporter, the lipid A transporter MsbA from Escherichia coli at 4.5 Å, representing the open

conformation [38], and the subsequent appearance of the structure of the transporter from Vibrio cholerae at 3.8 Å, showing the closed state [39], as well as the structure of the MsbA transporter from Salmonella typhimurium, which resembled the post-hydrolytic conformation [40], seemed to provide versatile starting points for protein homology modeling of ABCB1 in various states. However, the structures had to be withdrawn due to errors in data post-processing [41]. In 2006, two structures of a putative multi-drug transporter from methicillin-resistant Staphylococcus aureus (Sav1866) [42] [43] have been published, and subsequently, in November 2007, the corrected structures of MsbA [44] have been presented. Especially the structure of Sav1866, which exhibits a higherresolution (3.0 Å) than the new MsbA structures (5.5 Å), has been immediately used as template for the generation of protein homology models by Globisch et al. [45] and Ravna et al. [46] [47] to construct models of ABCB1 and ABCC5. The homology model of Globish et al. was subsequently used for the identification of putative binding sites for ligands. The authors were able to visualize multiple binding sites, which are mostly located in the transmembrane region of the model. However, Sav1866 has been crystallized in the post-hydrolytic state, which might limit the usability of the structural appearance of the identified residues. Furthermore, we were recently able to present a SAV1866-based homology model that was refined in a data driven structural modification approach to fit a vast collection of cross-linking data [48].

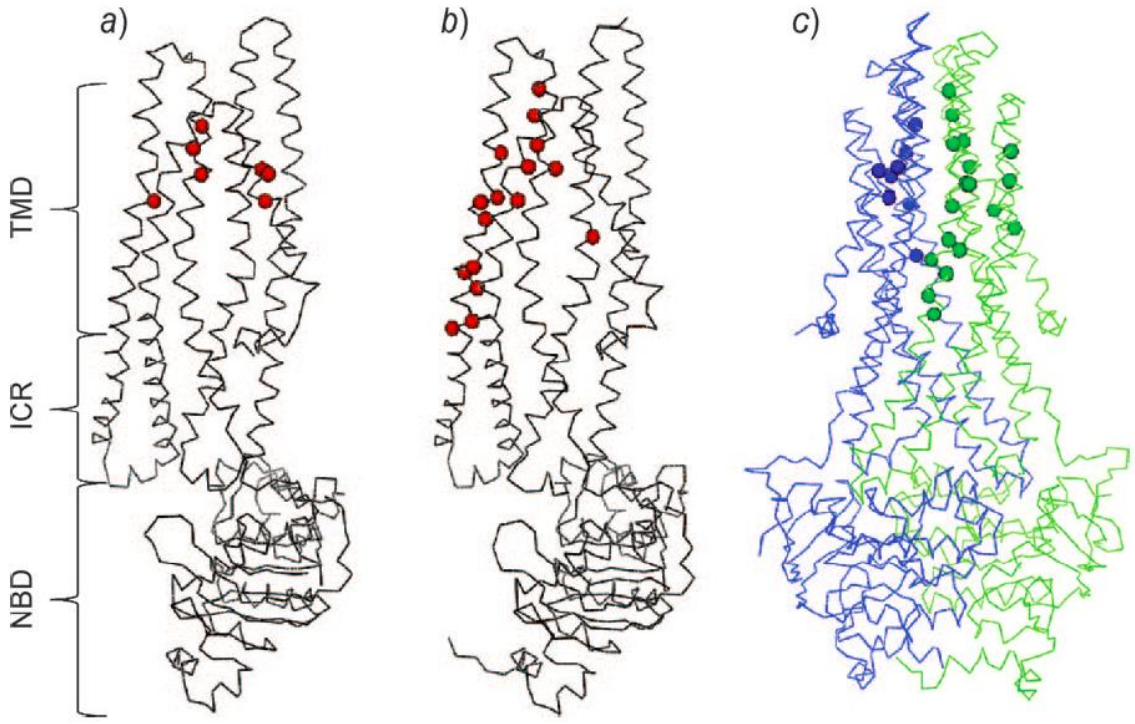


Figure 2: Mapping of Cα-atoms of published mutations that affect drug transport in the TMDs onto a model of ABCB1 based on the Sav1866 template and modified to fit cross-linking data.

Furthermore, amino acids that have been experimentally shown to alter substrate specificity after mutation cluster in distinct regions of the TMD 1 (Fig. 2,a) and TMD 2 (Fig. 2,b) are localized at the TMD/TMD interface (Fig. 2, c). In January 2009, Becker et al. published homology models of ABCB1 in different catalytic states by additionally utilizing the updated MsbA template structures [49]. They also performed docking of four different ligands into the non-energized (nucleotide-free) model and were able to show that all four ligands exhibit

interactions with experimentally shown residues. These first attempts on structure-based modelling of ABCB1-drug interactions seemed to be very promising. However, in the fourth March issue 2009 of Science, a new X-ray structure showing murine ABCB1 in a drug-binding competent state was published by Aller et al. [50]. This new structure, which represents the first mammalian ABC transporter structure, resembles the transporter in an initial (inward-facing) stage of the transport cycle (Fig. 3).

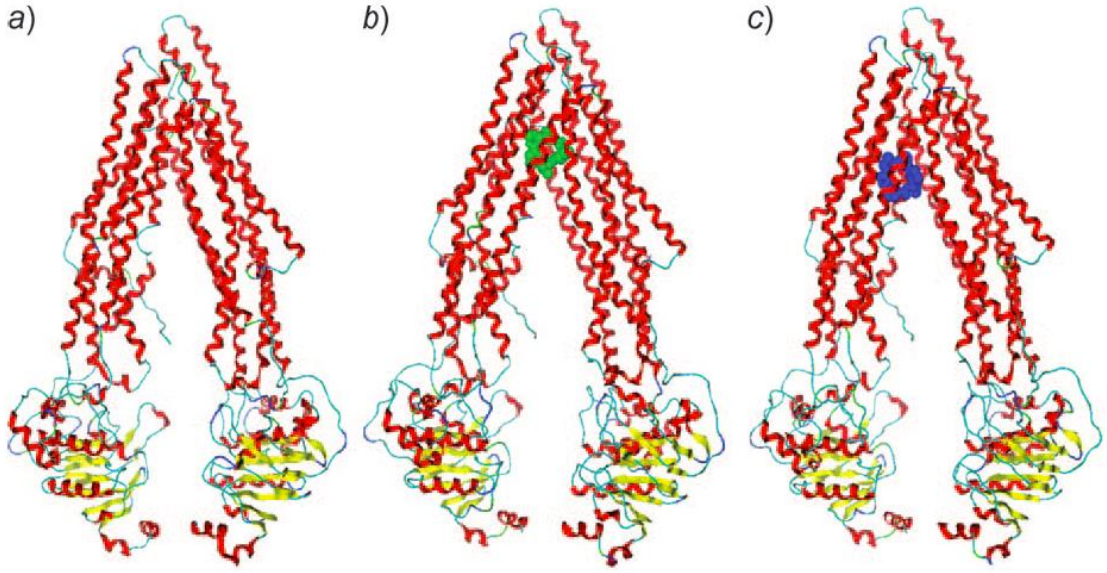


Figure 3: Recently published [50] crystal structures of murine ABCB1.

The murine ABCB1 structure (PDB code: 3g5u) has been crystallized at 3.8 Å resolution and shows a similar domain arrangement as the previously published bacterial homologues, SAV and MsbA. Additionally to this apo structure, the authors were able to crystallize drugbound ABCB1 (PDB code: 3G60, 3G61). Both, the apo as well as the drug-bound conformation, represent nucleotide-free states and show portals, spanned by TM segments, open to the cytoplasmic side and the inner leaflet of the plasma membrane. These new structures of ABC transporters definitely will have a large impact on the community and will strikingly affect prospective structure-based efforts.

Conclusions

In consequence of a steadily decreasing number of new chemical entities, the early prediction of toxicity, bioavailability, and potential drug – drug interactions is of utmost importance. In this respect, the important role of ABC transporters is increasingly recognized, and many pharmaceutical companies established high throughput assays for identifying ABCB1 substrates. However, utilizing the numerous data generated for creating accurate and highly predictive in silico models is still far from being optimal. Although most of the ligand-based studies show a satisfying performance, they lack interpretability which guides the medicinal chemist in the lead optimisation process. Furthermore, most of these models have been only evaluated in their study-specific, defined chemical space, but general applicability for the whole drug-like chemical space still needs to be proven. Thus, simple and widely applicable "rules of thumb" might represent useful alternatives to complement the more complex machine learning models. In any case, new studies using large diverse datasets, such

as those of Gleeson [32], have to be performed in the next future. The structure-based studies suffered for a long time from the disadvantage that the only template protein showing a non-energized, closed conformation was at rather low resolution. The recently published 3.8 Å X-ray structure of murine ABCB1, which shows 87% sequence identity to human ABCB1 and is additionally also co-crystallized with two inhibitors, provides a new and very promising starting point for docking and virtual screening attempts. Finally, one should always bear in mind that ABC transporters are part of a large metabolic network composed of nuclear receptors, cytochromes, and transporters. A complete understanding of the ADMET network, in which several ABC transporters play a crucial role, definitely requires a system-based view including and implementing in silico systems biology approaches.

We gratefully acknowledge financial support provided by the Austrian Research Promotion Agency (B1-812074) and by the Austrian Science Fund (F3502).

References

[1] H. Glavinas, P. Krajcsi, J. Cserepes, B. Sarkadi, Curr. Drug Delivery 2004, 1, 27.
[2] B. Hagenbuch, P. J. Meier, Pfluegers Arch. 2004, 447, 653.
[3] H. Koepsell, H. Endou, Pfluegers Arch. 2004, 447, 666.
[4] B. Sarkadi, L. Homolya, G. Szaka´cs, A. Va´ radi, Physiol. Rev. 2006, 86, 1179.
[5] M. M. Gottesman, T. Fojo, S. E. Bates, Nat. Rev. Cancer 2002, 2, 48.
[6] M. Dean, T. Fojo, S. E. Bates, Nat. Rev. Cancer 2005, 5, 275.
[7] G. Szakacs, A. Varadi, C. _zvegy-Laczka, B. Sarkadi, Drug Discovery Today 2008, 13, 379.
[8] R. L. Juliano, V. Ling, Biochim. Biophys. Acta, Biomembr. 1976, 455, 152.
[9] I. Nobeli, A. D. Favia, J. M. Thornton, Nat. Biotechnol. 2009, 27, 157.
[10] R. Didziapetris, P. Japertas, A. Avdeef, A. Petrauskas, J. Drug Targeting 2003, 11, 391.
[11] H. L. Pearce, A. R. Safa, N. J. Bach, M. A. Winter, M. C. Cirtain,W. T. Beck, Proc. Natl. Acad. Sci. U.S.A. 1989, 86, 5128.
[12] H. L. Pearce, M. A. Winter, W. T. Beck, Adv. Enzyme Regul. 1990, 30, 357.
[13] G. Ecker, Chemistry Today 2005, 23, 39.
[14] S. Ekins, R. B. Kim, B. F. Leake, A. H. Dantzig, E. G. Schuetz, L.-B. Lan, K. Yasuda, R. L. Shepard, M. A. Winter, J. D. Schuetz, J. H. Wikel, S. A. Wrighton, Mol. Pharmacol. 2002, 61, 974.
[15] S. Ekins, R. B. Kim, B. F. Leake, A. H. Dantzig, E. G. Schuetz, L.-B. Lan, K. Yasuda, R. L. Shepard, M. A. Winter, J. D. Schuetz, J. H. Wikel, S. A. Wrighton, Mol. Pharmacol. 2002, 61, 964.
[16] I. K. Pajeva, M. Wiese, J. Med. Chem. 2002, 45, 5671.
[17] J. E. Penzotti, M. L. Lamb, E. Evensen, P. D. J. Grootenhuis, J. Med. Chem. 2002, 45, 1737.
[18] T. Langer, M. Eder, R. D. Hoffmann, P. Chiba, G. F. Ecker, Arch. Pharm. (Weinheim, Ger.) 2004, 337, 317.
[19] A. Garrigues, N. Loiseau, M. Delaforge, J. Ferte´, M. Garrigos, F. Andre´, S. Orlowski, Mol. Pharmacol. 2002, 62, 1288.

[20] G. Cianchetta, R.W. Singleton, M. Zhang,M.Wildgoose, D. Giesing, A. Fravolini, G. Cruciani, R. J. Vaz, J. Med. Chem. 2005, 48, 2927.
[21] C. Chang, P. M. Bahadduri, J. E. Polli, P.W. Swaan, S. Ekins, Drug Metab. Dispos. 2006, 34, 1976.
[22] W.-X. Li, L. Li, J. Eksterowicz, X. B. Ling, M. Cardozo, J. Chem. Inf. Model. 2007, 47, 2429.
[23] G. Wolber, T. Seidel, F. Bendix, T. Langer, Drug Discovery Today 2008, 13, 23.
[24] M. A. Demel, R. Schwaha, O. Kr_mer, P. Ettmayer, E. E. J. Haaksma, G. F. Ecker, Expert Opin. Drug Metab. Toxicol. 2008, 4, 1167.
[25] J. Cramer, S. Kopp, S. E. Bates, P. Chiba, G. F. Ecker, Chem. Med. Chem. 2007, 2, 1783.
[26] V. K. Gombar, J.W. Polli, J. E. Humphreys, S. A. Wring, S. C. Serabjit-Singh, J. Pharm. Sci. 2004, 93, 957.
[27] Y. Xue, C.W. Yap, L. Z. Sun, Z.W. Cao, J. F.Wang, Y. Z. Chen, J. Chem. Inf. Model. 2004, 44, 1497.
[28] J. Huang, G. Ma, I. Muhammad, Y. Cheng, J. Chem. Inf. Model. 2007, 47, 1638.
[29] Y.-H. Wang, Y. Li, S.-L.Yang, L. Yang, J. Chem. Inf. Model. 2005, 45, 750.
[30] H. Li, C.W. Yap, C. Y. Ung, Y. Xue, Z. R. Li, L. Y. Han, H. H. Lin, Y. Z. Chen, J. Pharm. Sci. 2007, 96, 2838.
[31] A. Seelig, Eur. J. Biochem. 1998, 251, 252.
[32] M. P. Gleeson, J. Med. Chem. 2008, 51, 817.
[33] J. H. Friedman, B. E. Popescu, _Predictive Learning via Rule Ensembles. Technical Report_, 2005; http://www-stat.stanford.edu/~jhf/ftp/RuleFit.pdf.
[34] M. A. Demel, A. G. K. Janecek, W. N. Gansterer, G. F. Ecker, QSAR Comb. Sci. 2009, 28, 1087.
[35] M. F. Rosenberg, R. Callaghan, R. C. Ford, C. F. Higgins, J. Biol. Chem. 1997, 272, 10685.
[36] M. F. Rosenberg, A. B. Kamis, R. Callaghan, C. F. Higgins, R. C. Ford, J. Biol. Chem. 2003, 278,
8294.
[37] M. F. Rosenberg, R. Callaghan, S. Modok, C. F. Higgins, R. C. Ford, J. Biol. Chem. 2005, 280, 2857.
[38] G. Chang, C. B. Roth, Science 2001, 293, 1793.
[39] G. Chang, J. Mol. Biol. 2003, 330, 419.
[40] C. L. Reyes, G. Chang, Science 2005, 308, 1028.
[41] G. Chang, C. B. Roth, C. L. Reyes, O. Pornillos, Y.-J. Chen, A. P. Chen, Science 2006, 314, 1875.
[42] R. J. P. Dawson, K. P. Locher, Nature 2006, 443, 180.
[43] R. J. P. Dawson, K. P. Locher, FEBS Lett. 2007, 581, 935.
[44] A. Ward, C. L. Reyes, J. Yu, C. B. Roth, G. Chang, Proc. Natl. Acad. Sci. U.S.A. 2007, 104, 19005.
[45] C. Globisch, I. K. Pajeva, M. Wiese, Chem. Med. Chem. 2008, 3, 280.
[46] A.W. Ravna, I. Sylte, G. Sager, Eur. J. Med. Chem. 2008, 43, 2557.

[47] A.W. Ravna, I. Sylte, G. Sager, Theor. Biol. Med. Model. 2007, 4, 33.
[48] T. Stockner, S. J. de Vries, A. M. J. J. Bonvin, G. F. Ecker, P. Chiba, FEBS J. 2009, 276, 964.
[49] J.-P. Becker, G. Depret, F. van Bambeke, P. M. Tulkens, M. Pre´ vost, BMC Struct. Biol. 2009, 9, 3.
[50] S. G. Aller, J. Yu, A. Ward, Y. Weng, S. Chittaboina, R. Zhuo, P. M. Harrell, Y. T. Trinh, Q. Zhang, I. L. Urbatsch, G. Chang, Science 2009, 323, 1718.

Author Contribution

This chapter was published in *Chemistry and Biodiversity* as

Demel M., Krämer O., Ettmayer P., Haaksma E., Ecker G.F.**Predicting Ligand-Interaction with ABC-Transporters in ADME.** *Chemistry and Biodiversity, 2009*, 6, 1960-1969.

The initial manuscript for publication was written by me and reviewed, edited and improved by all the co-authors. All the calculations presented in this article were done by me. The results were evaluated by me. Prof. Ecker supervised this project.

Supplementary Information

Supplementary Table ST1: RuleFit Classification Results in 10-fold CV for the merged (LIT+NCI) model

CV run	tp	tn	fp	fn	acc	sens	spec	pract	prinact
1	21	16	8	5	0.74	0.81	0.67	0.72	0.76
2	23	18	5	4	0.82	0.85	0.78	0.82	0.82
3	19	17	6	8	0.72	0.70	0.74	0.76	0.68
4	17	23	4	6	0.80	0.74	0.85	0.81	0.79
5	9	21	10	10	0.60	0.47	0.68	0.47	0.68
6	20	18	6	6	0.76	0.77	0.75	0.77	0.75
7	16	16	10	8	0.64	0.67	0.62	0.62	0.67
8	19	19	9	3	0.76	0.86	0.68	0.68	0.86
9	17	14	10	9	0.62	0.65	0.58	0.63	0.61
10	19	20	4	5	0.81	0.79	0.83	0.83	0.80
mean	18	18	7	6	0.73	0.73	0.72	0.71	0.74
SD	4	3	2	2	0.08	0.12	0.09	0.11	0.08

CV run=crossvalidation run; tp=true positive; tn=true negative; fp=false positive; fn=false negative; acc=overall accuracy ((TP+TN)/(TP+TN+FP+FN)); sens=sensitivity (TP/(TP+FN)); spec=specificity (TN/(TN+FP)); pract=precision on actives (TP/(TP+FP)); prinact=precision on inactives (TN/(TN+FN))

Addendum: Recent Advances in Structure-Based Design regarding ABCB1 ligands

In the context of the above mentioned published paper, it is necessary to also discuss two more recent articles that have been published in 2011. These two papers apply target-based methods, including homology modelling and docking. Both articles use the new murine ABCB1 crystal structure mentioned above as basis for their modelling efforts. The first paper was published by our group (1), whereas the second paper was published by Bikadi and colleagues (2).

In the first article, Klepsch and colleagues describe the molecular basis of propafenone-type inhibitor binding to ABCB1 using a MODELLER derived homology model. Docking was done using the GOLD program. GOLD is a widely used, contemporary docking suite that utilizes a genetic algorithm based docking strategy. In a very elegant manner Klepsch et al. additionally employed an exhaustive-sampling strategy to investigate the established pose space by incorporating information from the well-established ligand-based SAR of the propafenone-type inhibitors (for review see (3)). Hence, this paper is the first which incorporates ligand-based SAR information into structure-based approaches to unravel the molecular basis of ABCB1 ligand binding. In this study, the predicted binding mode of a certain propafenone-derivative GPV062 showed hydrogen bonding interactions between the OH-group of this ligand and amino acid Y310. This docking finding is also in strong agreement with previously established SAR studies.

The second paper was presented by Bikadi and colleagues later in 2011 and represents a dual approach to predict ABCB1 mediate drug transport. The authors present a ligand-based SVM model based on 197 substrates and non-substres derived from the current literature and also docking into a homology model of the murine ABCB1 structure. The SVM model was validated using an external set of 32 compounds and showed a total accuracy of 75% and a MCC of 0.51. Furthermore, the authors also present a homology model of human ABCB1, which is subsequently subjected to docking. The presented docking results of rhodamine 123 were in good agreement with previously established biochemical data. A very important feature of the Bikadi publication is definitely that the authors made their models (both the SVM and docking model) publically available on a web-server (http://pgp.althotas.com).

The public availability of the docking model urged me to dock the propafenone GPV062 of the Klepsch study into the Bikadi model. This online tool is able to reliably predict the binding region of the fluorescent dye rhodamine 123 as well as that of digoxin. The result of the digoxin docking into the homology model of human ABCB1 is shown in the figure below.

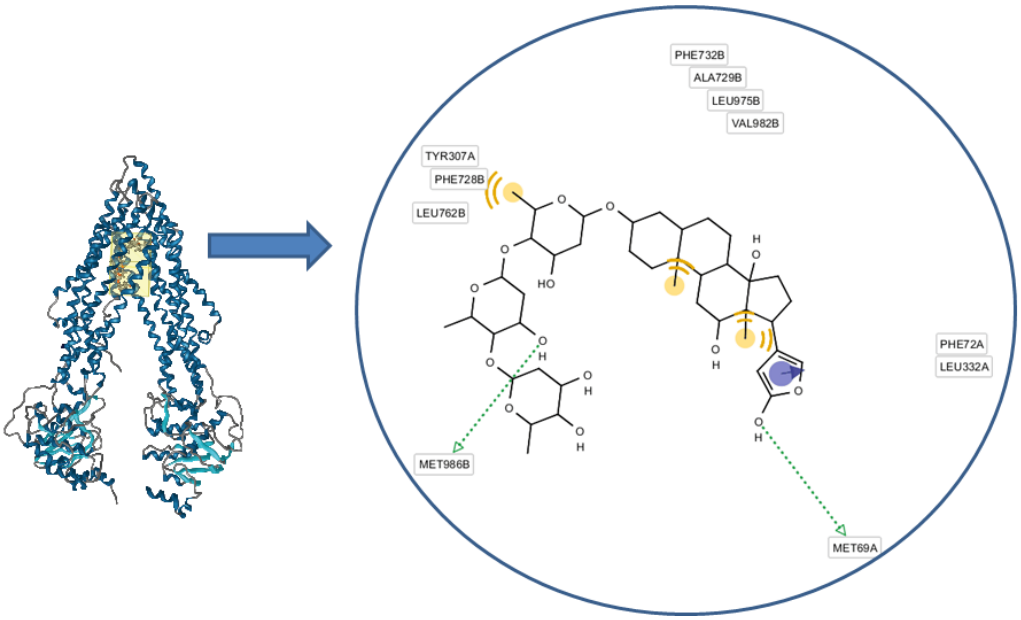


Figure 1: Docking of the „probe“ substrate digoxin into the ABCB1 homology model published by Bikadi et al. Docking was done using the automated web-server: http://pgp.althotas.com/. The graphic was renderd using LigandScout 3.03b evaluation version.

Despite this encouraging result, several ambiguities are encountered when the propafenone-type inhibitor GPV062 is docked into the homology model of Bikadi provided on the above mentioned web-server. The estimated binding energy of GPV062 is calculated by the web-server tool to be -10.9 kcal/mol. Notably, the retrieved result does not show any interaction with amino acid Y310 of the protein and the OH-moiety of GPV062, although this was predicted to be essential by SAR studies and was although further confirmed by the complex pose sampling strategy employed by Klepsch et al. The retrieved docking pose is also visualized in the figure below.

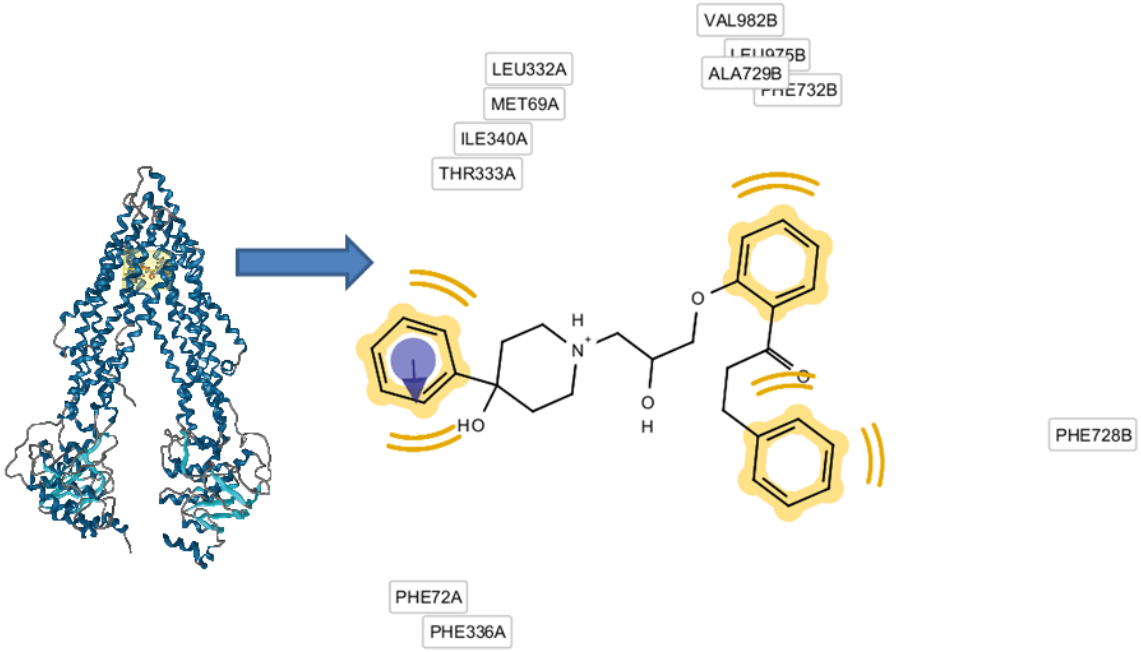


Figure 2: Docking of the propafenone-derivative GPV062 into the ABCB1 homology model published by Bikadi et al. Docking was done using the automated web-server: http://pgp.althotas.com/. The graphic was renderd using LigandScout 3.03b evaluation version. Notably, no interaction between the –OH group and the protein is observed.

The ambiguous results from the example shown above clearly highlight, that the analysis and use of structure-based approaches concerning ABCB1 must still be used with extreme caution. First of all it must be considered that structure-based approaches still harbour the disadvantage that homology models must be used as input for docking studies. Furthermore, if one takes into account that ABCB1 accomodates multiple binding sites incorporated into one huge binding cavity and considers that ABCB1 might be able to bind more than one ligand simultaneously, the uncertainty provided by the two different models and docking strategies, can be easily explained. This further suggests that strategies that utilize and incorporate previous SAR results in the pose selection process as done by Klepsch et al. might be in general more reliable than the use of binding-energy based scoring functions for analysing docking results into ABCB1 models. Hence it should be concluded that, despite the tremendous role of the murine ABCB1 structure, ligand-based approaches still represent the gold-standard in characterising ABCB1-mediated drug transport.

Bibliography

1. Klepsch F, Chiba P, Ecker GF. Exhaustive sampling of docking poses reveals binding hypotheses for propafenone type inhibitors of P-glycoprotein. PLoS Comput. Biol. 2011 Mai;7(5):e1002036.

2. Bikadi Z, Hazai I, Malik D, Jemnitz K, Veres Z, Hari P, u. a. Predicting P-glycoprotein-mediated drug transport based on support vector machine and three-dimensional crystal structure of P-glycoprotein. PLoS ONE. 2011;6(10):e25815.

3. Pleban K, Ecker GF. Inhibitors of p-glycoprotein--lead identification and optimisation. Mini Rev Med Chem. 2005 Feb;5(2):153–63.

Chapter 11: Publication III - Comparison of Contemporary Feature Selection Algorithms: Application to the Classification of ABC-Transporter Substrates

Abstract

Multidrug ABC-transporters are highly polyspecific in their substrate recognition pattern and influence the pharmacokinetics of a broad variety of structurally diverse compounds. Thus, prediction of ABC-transporter substrate properties of compound libraries is of major interest. In this study, we use k-nearest neighbor (kNN) classification in combination with five different feature subset selection (FS) algorithms to create predictive models for classification of ABCB1, ABCC1, and ABCG2 substrates. Our results show that FS methods that incorporate the classification algorithm give the best results and contain only a small subset of descriptors. For ABCB1 and ABCG2 cross validated accuracies of higher than 80% were achieved. The interpretation of the best performing feature subsets showed that descriptors consisting of simple counts as well as of projections of physicochemical properties on subdivided surface areas have highest contribution to the models.

Introduction

ATP-Binding-Cassette (ABC) transporters represent a ubiquitous family of membrane-bound proteins being mainly responsible for conducting chemo-defence mechanisms by extruding xenobiotics out of living cells [1]. Thus, the ABC-transporters ABCB1 (P-glycoprotein), ABCG2 (MXR, BCRP), and ABCC1 (MRP1) confer a multidrugresistant phenotype to cancer cells [2]. Furthermore, they are expressed in various tissues and thus influence absorption and distribution of a broad variety of structurally and functionally unrelated compounds [3]. In light of this increasing knowledge on the importance of ABC-transporters for bioavailability of candidate compounds prediction of potential substrates is of major interest in the early drug discovery phase [4]. Selecting the most relevant descriptors (features) reflecting the relationship between chemical structure and biological activity is one of the striking challenges in ligand-based design [5]. Roughly, feature subset selection (FS) algorithms can be categorized into two distinct classes: filters and wrappers. Filter methods are fast and classifier-agnostic, i.e. they do not rely on the performance of a specific classifier. Some of the filter methods consider interaction effects among variables and therefore return a selected feature subset, whereas others perform only feature ranking according to the individual predictive power of the respective descriptor. For ranking methods, an additional heuristic (e.g. selection of the n-top ranked features) has to be performed to yield a final subset. Wrappers are feedback methods, which rely on a specific classifier to evaluate the quality of a set of features. Thus, wrappers can also be seen as a feature subset selection method. Previously we were able to show that a different kind of data pre-processing technique, namely principal component analysis (PCA) can be successfully used as dimensionality reduction technique to classification of ABCB1 substrates by computing linear combinations of the original attributes and using them for classification. However, our results showed that classification performance with PCs is highly unstable and depends heavily on the methodologyapplied to calculate eigenvalues and eigenvectors [6]. Moreover, PC loadings do not always reliably indicate which variables are the most relevant ones [7, 8]. In this study we examine the performance of different FS methods that perform

either feature subset selection or feature ranking, rather than reducing the input space by linear combinations of these.

Methods

Data Sets

Data sets for ABCB1, ABCC1, and ABCG2 substrates have been obtained from the work of Szarkacs et al. [9]. Potential advantages of using these data for constructing in-silico models over certain other data sets (mostly compiledfrom different literature sources) are that there are no inter-laboratory differences in the measurements and it provides chemical information and activity of more than 1400 compounds for all 48 human ABC-transporters. These screening data contain Pearson`s correlation coefficients of the mRNA levels of the respective transporter and the cytotoxicity of a compound over a panel of 60 tumor cell lines In other words, compounds which give a negative correlation over 60 different cancer cell lines between transporter expression (determined as mRNA level) and their "intrinsic" cytotoxicity (the higher the expression of the transporter the lower is the toxicity of the compound) can be regarded as being transported by the transporter, whereas compounds showing no correlation between these two parameters are not regarded to interact with the protein. In this study we assign compounds with correlation coefficients lower than ≤ 0.3 to be substrates and those which show no correlation between toxicity and transporter expression ($-0.02<r<0.02$) to be nonsubstrates. This procedure yields a set of 240 (110 (45.8%) substrates and 130 (54.2%) nonsubstrates) compounds for ABCB1, 227 (124 (54.6%) substrates and 103 (45.4%) nonsubstrates) compounds for ABCC1 and 198 (94 (47.5%)substrates and 104 (52.5%) nonsubstrates) compounds for ABCG2.

Structure Preparation and Molecular Descriptors

Chemical structures were cleaned from counter ions, and hydrogens and lonepairs are added using the MOE2007.09 [10] wash routine. PEOE partial charges are assigned to each structure. Minimization was carried out using the MMFF99x forcefield. The descriptor classes used in this study cover the collection of all available 2D descriptors contained in the MOE software environment. We have discarded those descriptors that reflect only filter types (e.g.: number of leadlike-violations). In total we utilize 179 descriptors.

FS Algorithms

The five different FS techniques considered here embody one unsupervised subset selection method (RM), two filter ranking procedures (IG, ReliefF), one filter subset selection technique (CFS), and one wrapper for the kNN classifier (kNN-wrapper).

RM: The RM criterion, as proposed by McCabe [11, 12], is a proximity (or similarity) measure that uses the concept of matrix correlation of the PC matrix and the p-optimal feature subset of the original matrix. Matrix correlation is defined as the cosine between two $n \times p$ matrices [12]. This cosine origins from dividing the inner product of two matrices (which is calculated similarly to the "usual" inner product of two vectors) by the norm induced by this inner product. According to this, the RM coefficient represents the cosine of the PC matrix of the original matrix and a selected feature subset of the original data matrix. This feature

subset is selected by a genetic algorithm as search heuristic. The genetic algorithm utilizes the RM criterion as fitness function to identify the global optimal subset [13].

Information Gain (IG): IG, also known as Kullback-Leibler divergence, originally used to compute splitting criteria in decision tree algorithms, is often applied to find out how well each single feature separates a data set [14]. It can be seen as a supervised analogue of Claude Shannons information theoretical entropy calculation. The relevance of each attribute is measured in terms of entropy reduction. The underlying theory of this algorithm is to eliminate those descriptors whose value distributions are relatively random across the class labels, i.e., have only a small entropy. A drawback of this filter-ranking method is that each descriptor is evaluated independently of the context of other descriptors.

ReliefF: The ReliefF algorithm, as introduced by Kira and Rendell [15, 16], is a ranking method which, utilizes instance based learning to determine a relevance weight for each descriptor. Each of these weights reflect the descriptors ability to discriminate between the classes. The output of this method is a ranked quality weight for each feature in the range [-1,+1].

Correlation-based Feature Selection (CFS): Contrary to IG and ReliefF, CFS as introduced by Hall [17] is a filter method that performs feature subset selection. The resulting feature subset contains features that show a high degree of correlation with the class label (i.e., they are supposed to be predictive of the class label), while having a low degree of intercorrelation (i.e., are not supposed to be predictive of each other).

kNN-Wrapper: Contrary to the methods above, wrappers incorporate the machine learning algorithm in the FS process, rather than being independent of it [18]. Wrappers, generally, tend to achieve better classification results than filter methods based on the fact that they are tuned towards the classification algorithm and its training data. However, they tend to be much slower than filters because they must repeatedly call the induction algorithm. In our case, the resulting feature subset from a wrapper relies on the performance of the kNN algorithm evaluated by 10-fold cross validation.

Classification Algorithm

For comparing the effectiveness of our FS methods, we apply kNN classification modeling. A 10-fold cross validation is carried out to determine the performance of the different models. Each data set containing the features selected by the respective FS method is randomly split into 90% training compounds and 10% test compounds. A kNN model is constructed on the selected 90% training compounds and the 10% test compounds are predicted. This is repeated 10 times. It is noteworthy to mention, that sampling is done without replacement in order to assure that each compound is one time in the test set and the remaining nine times in the training sets. Additionally, we used Y-randomization to estimate the relationship between the derived feature subsets and the binary biological activity. For this, we randomly split our data sets into 90% training and 10% test set and permuted the

class label of the training set. Consecutively, a kNN model was built on this data set of the selected features for each method and this permuted activity. This procedure was necessary since kNN modelling has no intrinsic training step (kNN is a "lazy" learning method or "instance-based" learning method, which does classification in the same step as learning). Afterwards, this model was used to predict the class label of the remaining 10% test compounds. This was repeated for 100 times and classification results were averaged for each method and each data set. The overall classification accuracy (given in percent) was calculated as: ((true positives+true negatives) / all compounds)*100.

Software

The R software package [19] was used to generate the RM subsets (function genetic in the subselect package) as well as for generating the 2D radial visualization plots (function radviz2d in the dprep package). The Y-randomized training sets for each data set were generated using an inhouse R script. All other FS methods as well as the classification were performed using the WEKA3.5.8 software [20].

Results and Discussion

Comparison of FS Methods

We compared five different FS algorithms with respect to their classification accuracy in 10-fold cross validation runs as well as in 100 replicates of Y-randomization. Furthermore, we compared the derived models also in terms of the numbers of descriptors. Since classification was done on basis of kNN we first elucidated the optimal value of k. For all data sets it is shown that k=1 retrieved the best results (see Supplementary Information SI 1). Therefore, the results in Table 1 report classification with k=1. From Table 1 it can be seen that the best models in terms of %-classification accuracy are derived by the wrapper FS methodology with correct classifications of 85.6% for ABCB1, 72.0% for ABCC1, and 88.1% for ABCG2 in 10-fold crossvalidaton. Comparing these kNN classification results for ABCB1 with previously published results with kNN learning on higher dimensional data sets [21] demonstrates the applicability of our models. Our models also show a relatively low classification accuracy in Y-randomization, which further highlights the information content of the selected features. Additionally, the models contain only a small number of descriptors (10 – 12) for all three targets. The unsupervised method (RM) showed the poorest classification performance for the three data sets. The ranking methods IG and ReliefF gave similar classification results, but the best models retrieved with the ReliefF method contained a smaller number of descriptors (see Supplementary Information SI2). However, ReliefF also showed a similar performance for Y-randomization as for cross-validation (especially for ABCC1), which renders this method questionable for this application. The CFS algorithm showed a medium performance among the FS methods used. Since the best models were obtained when applying the *k*NN-wrapper algorithm, we also examined the classification performance of the three feature sets with respect to other classification algorithms such as support vector machine and decision tree.

Table 1: Classification performance expressed as classification accuracy for the three data sets and the five FS methods. Best classification results are shown in bold letters, worst results are underlined; ACC¼classification accuracy, 10xCV= 10-fold cross-validation, Yrand=Y-randomization, #descr=number of descriptors.

	ABCB1			ABCC1			ABCG2		
	ACC(10xCV)	Yrand(test)	# descr	ACC(10xCV)	Yrand(test)	# descr	ACC(10xCV)	Yrand(test)	# descr
RM	63.54	54.55	7	57.66	34.78	9	44.76	55.22	2
IG	75.42	59.09	90	62.11	47.83	54	76.91	50.69	72
RELIEF	80.83	72.73	54	65.11	65.22	18	76.56	55.36	18
CFS	74.6	36.36	15	59.44	52.17	3	79.07	70.69	11
WRAPPER	**85.57**	45.46	12	**71.96**	39.13	10	**88.14**	64.55	11

For the ABCB1 and the ABCG2 data set classification accuracies of 75% to 79% are obtained. For the ABCC1 set classification results are worse, but still comparable to the *k*NN results. This suggests that the obtained feature sets might also be used successfully in the context of other machine learning systems. For details see Supplementary Information SI3.

Interpretation of Wrapper Selected Feature Subsets

For the three ABC-transporters the following descriptors have been selected by the wrapper method:

- ***ABCB1:*** apol, chi0_C, chi0v_C, chi1_C, rings, PEOE_VSA-5, PEOE_VSA_POL, PEOE_VSA PPOS, SlogP_VSA0, SMR_VSA2, TPSA, opr_brigid.
- **ABCC1:** a_count, a_hyd, chi1v, opr_nring, PEOE_VSAþ3, PEOE_VSAþ5, PEOE_VSA-4, PEOE_VSA-6, Q_VSA_PNEG, vsa_acc.
- **ABCG2:** a_count, a_hyd, a_nC, a_nH, chi1v, SlogP_VSA1, SlogP_VSA2, SlogP_VSA8,
  SMR_VSA1, SMR_VSA6, VDistMa.

From this list of selected descriptors it can be deduced that simple atom counts as well as a description in terms of subdivided surface areas (VSA) are in general good means to describe substrates and nonsubstrates of ABC-transporters. Interestingly, for ABCB1 and ABCG2 VSA-descriptors reflecting lipophilicity and size (SlogP_VSA and SMR_VSA) are selected, whereas for ABCC1 the dominant class of VSA-descriptors is the partial charge class (PEOE_VSA). The three descriptor sets are graphically visualized in a 2D-radial visualization in Figure 1A–C, which represents a nonlinear projection of the attribute space with the attributes shown along the perimeter of the circle onto two dimensions. Classes are specified by circles (active/substrates) and squares (inactive/nonsubstrates). For details on this visualization technique see [22]. The graphs show that for ABCB1 and ABCG2 a good separation of the two classes is achieved using the wrapper selected attributes. Maximum

class separation is considered to be one of the striking attributes of a good feature subset. The results in Figure 1A –C illustrate the class discriminating power of the selected subsets. The interpretation of Figure 1A highlights that substrates of ABCB1 have higher values for SMR_VSA2 and SlogP_VSA8 and the nonsubstrates show a higher number of rigid bonds (Oprea_srigid bond count – opr_brigid), polarizable atoms (a_pol) and higher values for partial charge descriptors (PEOE_VSAPPOS, PEOE_VSA_POL). The importance of lipophilicity for ABCB1 substrates has already been mentioned in other studies [23, 24]. From these results, a reduction of lipophilicity as well as an increase in the number of rigid bonds might be a promising strategy to avoid ABCB1 conferred drug-transport. For ABCG2 (Fig. 1C) SMR_VSA1 and SMR_VSA6 are the dominant descriptors for the active class, whereas the inactive class shows higher values for SlogP_VSA1, SlogP_VSA2, and a_hyd. For the ABCC1 set (Fig. 1B) the separation is only moderate, which is in convergence with its weak classification performance. However, the nonsubstrates seem to be more rigid, which is reflected in Oprea_s ring count descriptor.

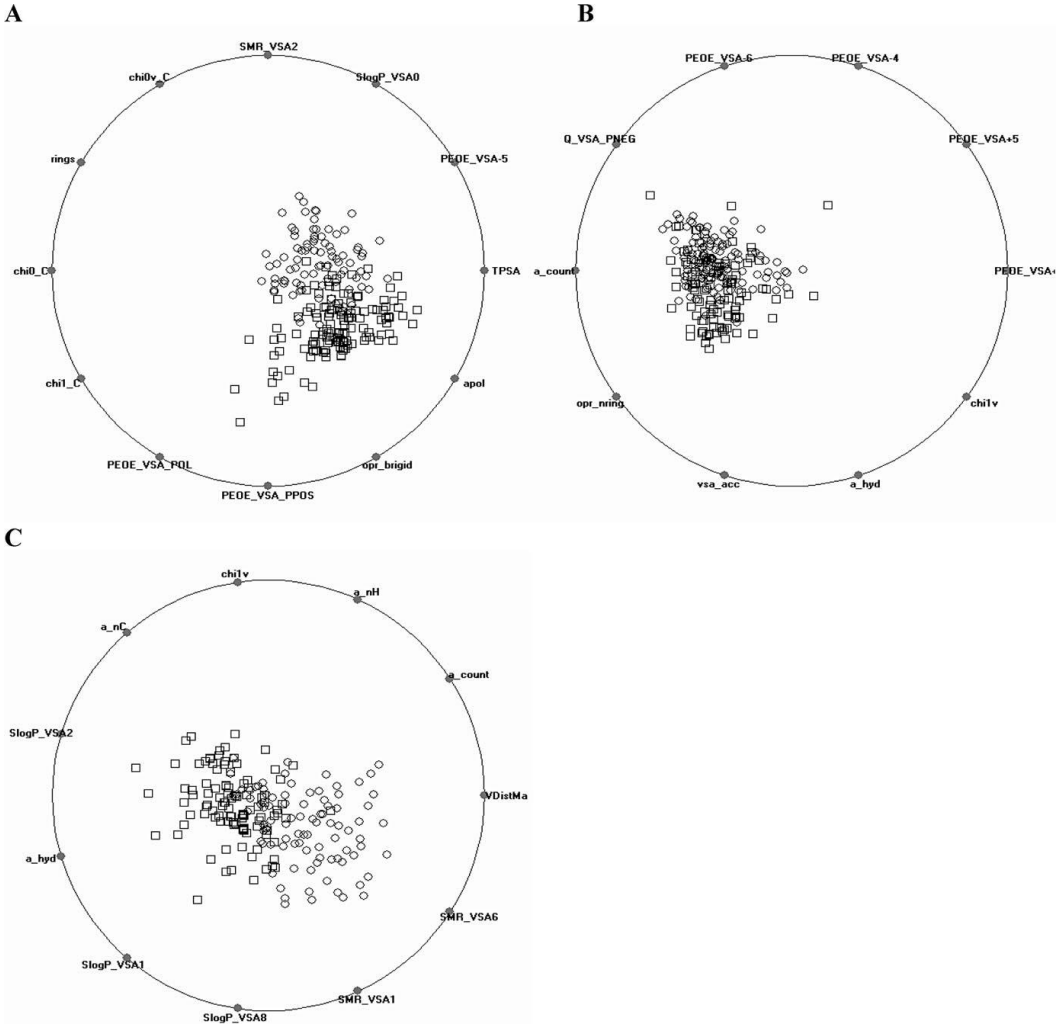


Figure 1: A– C: Radial visualization of the kNN-WRAPPER selected descriptors. Classes are encoded as follows: circles=substrates/actives, squares=nonsubstrates.

Conclusion

In this paper we concentrated on the classification performance of five different feature subsets for the three ABCtransporters ABCB1, ABCC1, and ABCG2. Our results show that

the wrapper method outperforms the other FS methods. Additionally, a comparison of the three feature sets retrieved for ABCB1, ABCC1 and ABCG2 highlights certain general properties (e.g. size, partial charge, rigidity) of ABC-transporter substrates and nonsubstrates that might be useful in shaping chemical libraries to avoid ABC-transporter related ADMET problems.

Acknowledgements

This work was supported by the FFG (grant #B1-812074) as well as by the CPAMMS project (FS397001) in the research focus area "Computational Science" of the University of Vienna.

References

[1] K. Linton, Physiology 2007, 22, 122 – 130.

[2] M. M. Gottesman, T. Fojo, S. E. Bates, Nat. Rev. Cancer 2002, 2, 48 – 58.

[3] G. Szaka´ cs, A. Va´ radi, C. _zvegy-Laczka, B. Sarkadi, Drug Discov. Today 2008, 13, 379 – 393.

[4] G. Ecker, Chemistry Today 2005, 23, 39 – 42.

[5] M. A. Demel, A. G. K. Janecek, K.-M. Thai, G. Ecker, W. N. Gansterer, Curr. Comput.-Aided Drug Design 2008, 4, 91 – 110.

[6] A. G. K. Janecek, W. N. Gansterer, M. A. Demel, G. F. Ecker, JMLR 2008, 4, 90 – 105.

[7] J. F. C. L. Cadima, I. T. Jolliffe, J Appl. Statist. 1995, 22, 203 – 214.

[8] J. F. C. L. Cadima, I. T. Jolliffe, J Agric. Biol. Environ. Statist. 2001, 6, 62 – 79.

[9] G. Szakacs, J. P. Annereau, S. Lababidi, U. Shankavaram, A. Arciello, K. J. Bussey, W. Reinhold, Y. Guo, G. D. Kruh, M. Reimers, J. N. Weinstein, M. M. Gottesman, Cancer Cell 2004, 6, 129 – 137.

[10] Molecular Operating Environment (MOE), Chemical Computing Group, Version 2007.09.

[11] G. P. McCabe, Technical Report #86 – 19, Dept. of Statistics, Purdue University 1986.

[12] G. P. McCabe, Technometrics 1984, 26, 137 – 144.

[13] J. F. C. L. Cadima, I. T. Jolliffe, Comput. Stat. Data Anal. 2004, 47, 225 – 236.

[14] I. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, San Francisco, CA 2005.

[15] K. Kira, L. Rendell, in Ninth International Workshop on Machine Learning 1992, 249 – 256.

[16] I. Kononenko, in European Conference on Machine Learning, 1994, 171 – 182.

[17] M. Hall, PhD thesis, Waikato, New Zealand, 1998.

[18] R. Kohavi, G. H. John, Artificial Intelligence 1997, 97, 273 –324.

[19] http://cran.r-project.org/

[20] http://www.cs.waikato.ac.nz/_ml/weka/

[21] P. De Cerqueira-Lima, A. Golbraikh, S. Oloff, Y. Xiao, A. Tropsha, J. Chem. Inf. Model. 2006, 46, 1245 – 1254.

[22] M. Ankerst, D. Keim, in Proc. IEEE Visualization Conf. 1997.

[23] R. Didziapetris, P. Japertas, A. Avdeef, A. Petrauskas, J. Drug Targeting 2003, 391 – 406.

[24] J. Huang, G. Ma, I. Muhammad, Y. Cheng, J. Chem. Inf. Model. 2007, 47, 1638 – 1647.

Author Contribution

This chapter was published in *QSAR and Combinatorial Sciences* as

Demel M., Janecek A.G.K., Gansterer W.N., Ecker G.F.**Comparison of Contemporary Feature Selection Algorithms: Application to the Classification of ABC-Transporter Substrates.***QSAR and Combinatorial Sciences, 2009,* 28, 1087-1091.

Parts of this chapter were also presented at the EuroQSAR conference, 2008.

The initial manuscript for publication was written by me and reviewed, edited and improved by all the co-authors. PD Dr. Gansterer and Dr. Janecek assisted in the design of experiments and supported me with their extensive experience in the field. All the calculations were done by me. The results were evaluated by me. Prof. Ecker supervised this project.

Supplementary information:

Supplementary Information SI 1: The dependency of k in kNN classification

PCA-GA & CFS & WRAPPER:

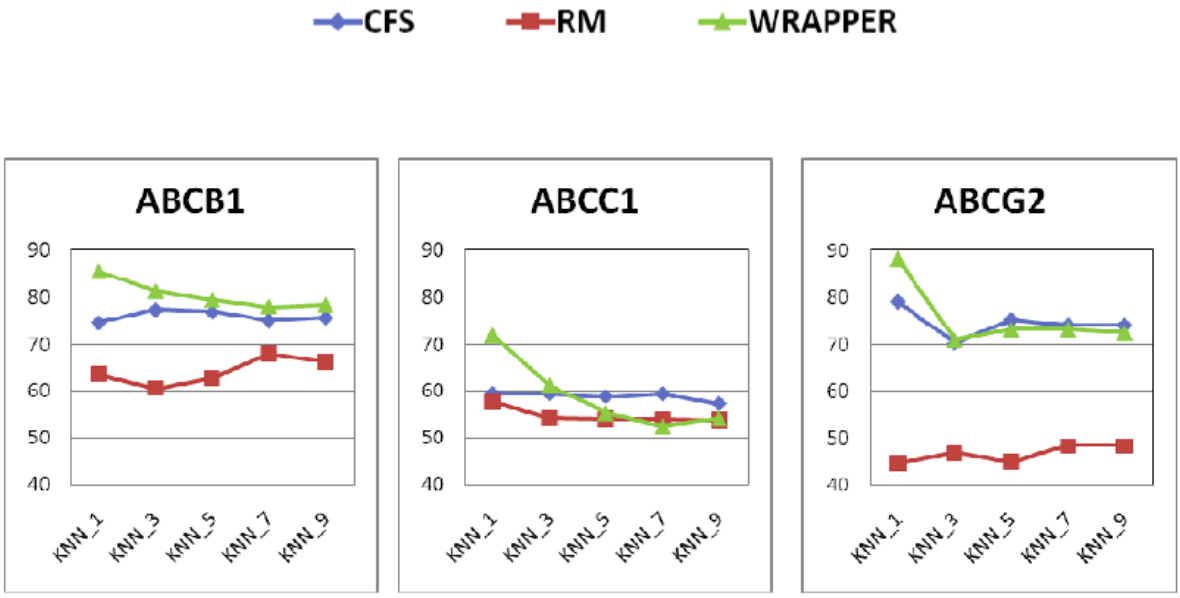


Fig. SI1.1 RM, CFS and WRAPPER. The x-axis denotes the different number of k neighbors and the y-axis shows the classification accuracy.

Information Gain:

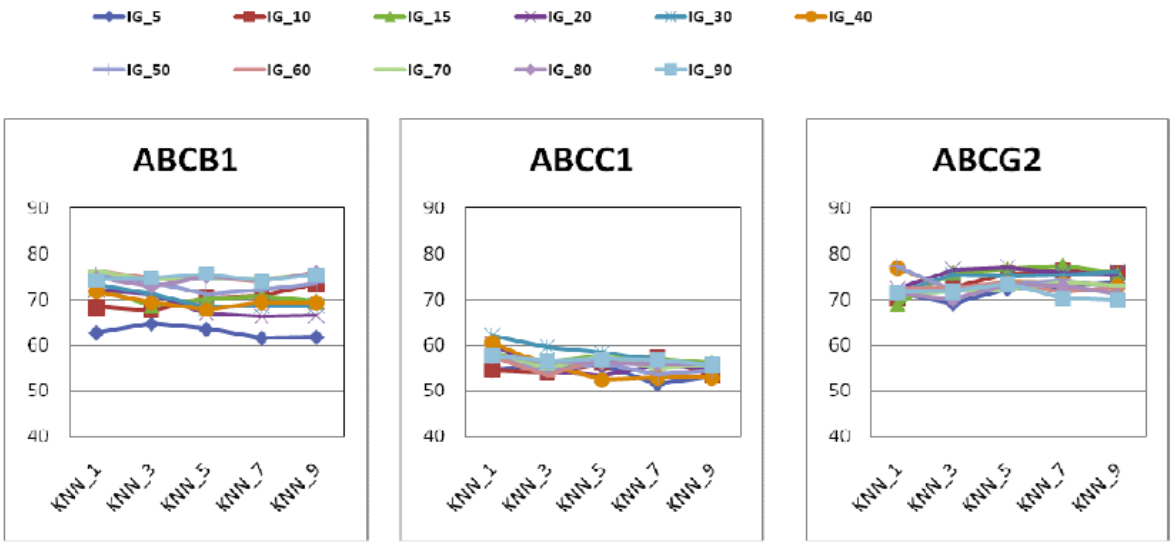


Fig. SI1.2 Information Gain. The x-axis denotes the different number of k neighbors and the y-axis shows the classification accuracy.

RELIEFF:

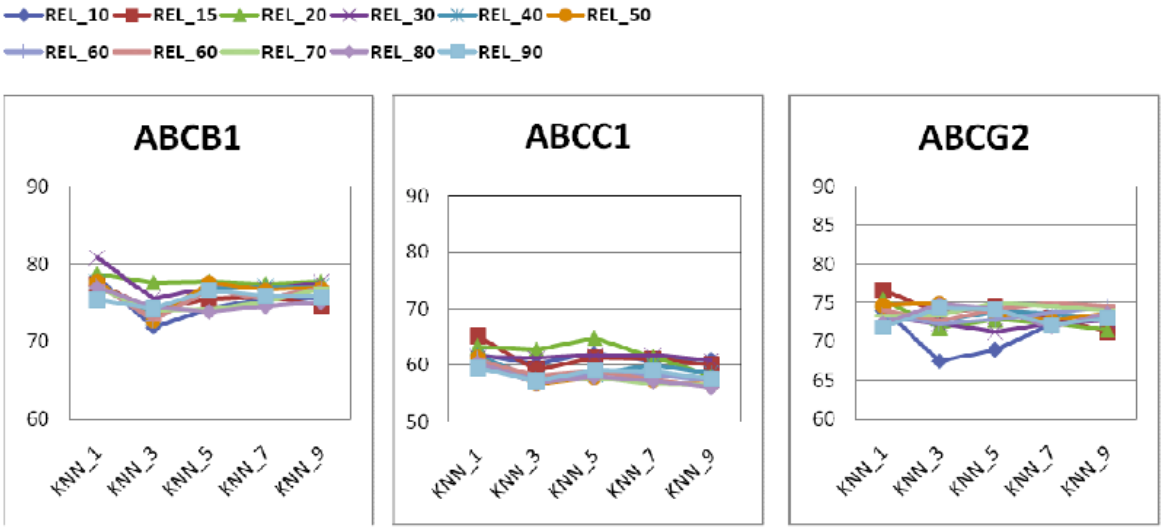


Fig. SI1.3 RELIEFF. The x-axis denotes the different number of k neighbors and the y-axis shows the classification accuracy.

Supplementary Information SI 2: Comparison of filter ranking methods – Information Gain vs. ReliefF

	IG_ABCB1	RELIEF_ABCB1	IG_ABCC1	RELIEF_ABCC1	IG_ABCG2	RELIEF_ABCG2
IG_ABCB1	1					
RELIEF_ABCB1	0.381	1				
IG_ABCC1	0.070	0.0818	1			
RELIEF_ABCC1	0.150	0.2242	0.002	1		
IG_ABCG2	0.607	0.1870	0.110	0.103	1	
RELIEF_ABCG2	0.076	0.3664	0.132	0.426	0.29913	1

Figure SI2.1 Correlation Coefficients between IG- and RELIEFF- scores across the three data sets. It can be seen that there is no good correlation between either the methods (IG and RELIEFF) or the three data sets. This supports the fact that both methods select different descriptors for the same data set and that there is no similar ranking between the three data sets.

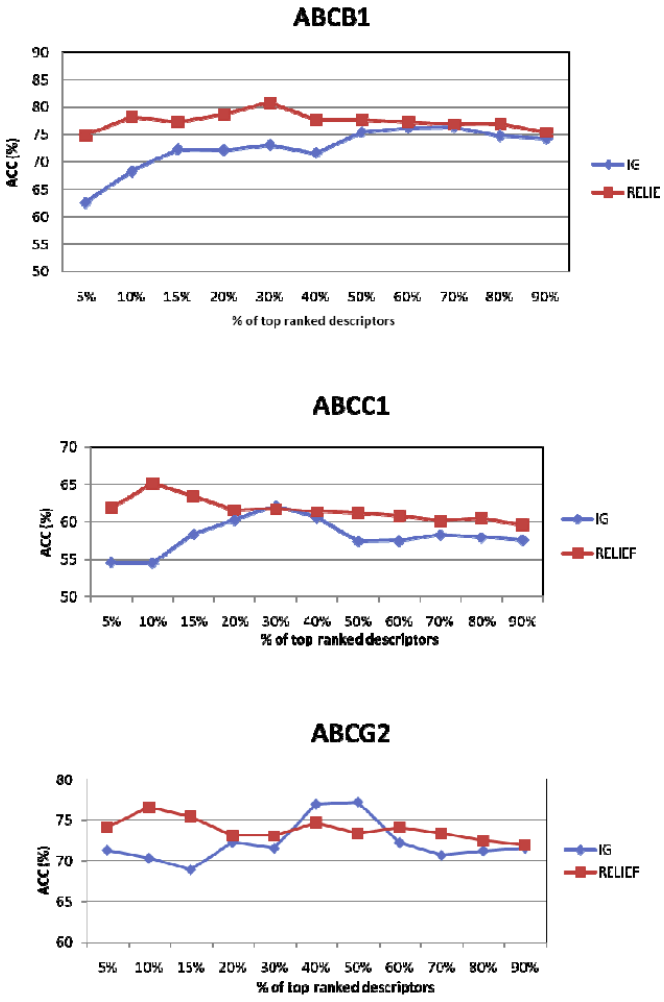


Figure SI2.2 Comparison of IG and RELIEFF for the three data sets. Results are shown in terms of 10xCV for the three data sets (A-C). IG performance is shown in blue and RELIEFF performance is depicted in red. The figures show that independently of the data set the performance of these two FS method is quite similar until a reduction of the input space of about 50% (e.g.: ABCB1) or even 20% (ABCG2) is reached. A further reduction of the feature space results in a decrease in classification performance for the IG selected

subsets. Contrary, to that at 10% to 15% of the top ranked features RELIEFF shows an increase in classification accuracy. This shows that independent of the data set, best classification results (among these two methods) aren received with very small descriptor sets, when using the RELIEF algorithm.

Supplementary Information SI 3:

Evaluation of the kNN-WRAPPER feature sets and the RM subset by other machine learning algorithms

DATASET	SET	kNN(k=1)	SVM	DT
ABCB1	WRAPPER	85.57	75.60	72.10
ABCC1	WRAPPER	71.96	65.68	64.69
ABCG2	WRAPPER	88.14	77.55	73.27

Figure SI3.1 Classification performance of the kNN-WRAPPER subsets, when classified by a SVM and DT.
The hyperparameters for the SVM: *polynomial kernel* for feature transformation and the "soft margin"-parameter *C* is set to 1.0; for the DT: J.48 algorithm with tree pruning enabled and a confidence factor of 0.25. The results of these classifications correspond to a 10-fold cross validation.

DATASET	SET	kNN(k=1)	SVM	DT
ABCB1	RM	63.54	55.20	56.72
ABCC1	RM	57.66	53.56	51.29
ABCG2	RM	44.76	47.96	44.89

Figure SI3.2 Classification performance of the RM subsets, when classified by a SVM and DT. The hyperparameters for the SVM: *polynomial kernel* for feature transformation and the "soft margin"-parameter *C* is set to 1.0; for the DT: J.48 algorithm with tree pruning enabled and a confidence factor of 0.25. The results of these classifications correspond to a 10-fold cross validation

Chapter 12: Publication IV - Ensemble Rule-Based Classification of Substrates of the Human ABC-Transporter ABCB1 Using Simple Physicochemical Descriptors

Abstract

Within the last decades, the detailed knowledgeon the impact of membrane bound drug efflux transporters of the ATP binding cassette (ABC) protein family on the pharmacological profile of drugs has enormously increased. Especially, ABCB1 (P-glycoprotein, P-gp, MDR1) has attracted particular interest in medicinal chemistry, since it determines the clinical efficacy, side effects and toxicity risks of drug candidates. Based on this, the development of in silico models that provide rapid and cost-effective screening tools for the classification of substrates and nonsubstrates of ABCB1 is an urgent need in contemporary ADMET profiling. A characteristic hallmark feature of this transporter is its polyspecific ligand recognition pattern. In this study we describe a method for classifying ABCB1 ligands in terms of simple, conjunctive rules (RuleFit) based on interpretable ADMET features. The retrieved results showed that models based on large, very diverse data sets gave better classification performance than models based on smaller, more homogenous training sets. The best model achieved gave a correct classification rate of 0.90 for an external validation set. Furthermore, from the interpretation of the best performing model it could be concluded that in comparison to nonsubstrates ABCB1 substrates generally show a higher number of hydrogen-bond acceptors, are more flexible and exhibit higher logP values.

Introduction

Almost one third of all compounds in the drug discovery pipeline fail due to improper ADMET (absorption, distribution, metabolism, elimination, toxicity) behaviour which renders in silico prediction of ADMET properties an important issue.[1] Within the past decade a variety of different transporter protein families have been shown to influence the pharmacokinetics of many drugs. ATP binding cassette (ABC) transporters, which are localized in the cell membrane, are known to confer a multidrug-resistant (MDR) phenotype to cancer cells by extruding chemotherapeutic agents out of the tumour cells.[2] However, the major physiological role of the ABC multidrug transporters is to protect cells and tissues against xenobiotics. Therefore, these transporters additionally play a crucial role in drug availability, metabolism and toxicity.[3,4] ABCB1 (P-glycoprotein, Pgp, MDR1) represents the paradigm transporter in the field and has been the focus of intense research since its discovery more than three decades ago. ABCB1 contributes to transporting a variety of chemical and pharmacological unrelated molecules.[5,6] This polyspecific substrate recognition pattern in conjunction with the absence of a high resolution structure of human ABCB1, renders ligand-based design strategies the most promising attempts to identify potential ABCB1 substrates early in the drug discovery process.

Substantial effort has been made to address the ABCB1-substrate classification problem. For instance, Seelig suggested that molecules can be categorized as ABCB1 substrates if they contain structural recognition patterns that are formed by two or three hydrogen bond acceptor groups that are either in a 2.5 Å or 4.6 Å fixed spatial separation.[7, 8] Didziapetris and co-workers postulated the “rule of four”, which states that compounds with the number of

hydrogen bond acceptors in a molecule (N+O)≥8, and a molecular weight (MW)>400 and most acidic pKa>4 are likely to be ABCB1 substrates whereas compounds with (N+O)≤4, MW<400, and most basic pKa<8 probably are nonsubstrates.[9] In 2004 Gombar et al. postulated the "Gombar-Polli Molecular E-state (MolES) Rule". Molecules with MolES>110 showed to be substrates and those with MolES<48 showed to be nonsubstrates.[10] More recently, Gleeson et al. suggested that neutral or basic molecules showing a MW>400 and/or a logP value>4 are more likely to be transported by ABCB1 than acidic or zwitterionic compounds.[11] Other authors used machine learning algorithms, such as support vector machines (SVM) and artificial neural networks, often in conjunction with highly complex descriptors, to classify ABCB1 substrates.[12–18] For instance, Xue and colleagues constructed an SVM model that gave an overall accuracy of 80%.[16] Cabrera et al. applied the TOPSMODE (topological substructural molecular design) approach to classify ligands of ABCB1.[19] Wang et al. used both supervised back propagation neural networks and unsupervised Kohonen self-organising maps (SOMs) to construct models that discriminate between ABCB1 substrates and inhibitors.[17] They used stepwise linear discriminant analysis to select 11 relevant descriptors out of a pool of 287 descriptors. The final model gave a sensitivity of 83.3% and a specificity of 80.8% using SOMs. De Cerqueira Lima et al. applied the combinatorial QSAR approach to an 195 compound ABCB1 substrate/nonsubstrate data set. Thereby, they compared the performance of different machine learning algorithms, such as k-nearest neighbor, decision tree, BinaryQSAR, and SVM, and different descriptor types, such as MolconnZ, atom pair descriptors, VolSurf and MOE 2D-descriptors.[18] The most successful approach so far has been introduced by Huang and colleagues in 2007.[12] They combined a particle swarm feature selection preprocessing algorithm with a subsequent SVM classification model and achieved an accuracy of 88.6% on an external test set of 40 compounds.

Although in general highly predictive, most of the algorithms do not allow a direct guidance of the medicinal chemist with respect to chemical features important for substrate properties. In this study we utilize Friedman's RuleFit algorithm, which allows the generation of an ensemble of conjunctive rules (i.e., two or more conditions are connected via logical AND) in form of a linear model, to classify ABCB1 substrates and nonsubstrates.[20] We especially aim to provide predictive and also interpretable machine learning models that both (i) allow the rapid in silico identification of potential ABCB1 substrates and (ii) provide the medicinal chemists with useful information on simple chemical features that might render new lead candidates ABCB1 substrates or nonsubstrates.

Material and Methods

Data Sets

In this work we utilize three data sets that are constituted out of chemical structures derived from different in-house hit-to-lead optimization projects at Boehringer Ingelheim. The ABCB1 substrate liability for all these cytotoxic compounds was assessed in an assay using the sensitivity of a transporter overexpressing cell line. Briefly, activities against the vincristine resistant and ABCB1-overexpressing cell line H460vinc were measured and compared to the activities against the wild-type cell line $H460_{wt}$. In total 3313 compounds

have been assayed under the same conditions using the same assay protocol. While all compounds display activity on the wild-type cell line (EC_{50} $H460_{wt} \leq 1$ mM), compounds with an ABCB1 substrate liability show decreased activity on the ABCB1-overexpressing cell line $H460_{vinc}$. In order to construct two-class classification models we annotated the biological activity of our molecules (+1 for substrates; -1 for nonsubstrates) according to the following thresholds:

Substrates: EC50 $H460_{vinc}$/EC50 $H460_{wt}$>45
Nonsubstrates: EC50 $H460_{vinc}$/EC50 $H460_{wt}$<5

A compound showing a ratio of EC_{50} values smaller than 5 was considered not to have ABCB1 substrate liability and is therefore annotated as a nonsubstrate. In order to account for errors typically associated with cytotoxicity assays, which are in the range of a factor of 2–3, compounds showing an intermediate ratio ($5 \leq EC_{50}$ $H460_{vinc}/EC_{50}$ $H460_{wt} \leq 45$) were not annotated and withdrawn from the data set. A potential advantage of these data sets in comparison to some other sets used in previous studies, is definitely that they do not suffer from differences in inter-laboratory measurements, however it needs to be stated that this kind of assay does not allow a discrimination of nonsubstrates and inhibitors (for specific comments on this, see also [6]). Three training data sets were used in this work for model generation. The first one represents a small, chemically very homogenous collection consisting of 227 molecules (HOM) from one hit-to-lead project, whereas the second data set represents a big and rather diverse collection of 1650 compounds (DIV). Whereas the HOM data set resembles an almost balanced class distribution, the DIV data set shows a probably more realistic, unbalanced class distribution (87% nonsubstrates, 13% substrates). The third training set represents the merged combination of the aforementioned two data sets and consists of 1877 molecules (MERGED). Furthermore, we generated an independent validation (EXT) set that consists of additional 1436 compounds. These compounds are from the same hit-to-lead projects as the compounds in the training sets, but have been synthesized and tested during the progress of this study. The ratio of substrates and nonsubstrates for the training sets as well as the external set are given in Table 1.

Table 1: Data sets used in this study.

	# sub[a]	# nonsub[b]	# total[c]	avg. similarity[d]
HOM	129 (56.8%)	98 (43.2%)	227	0.83 (±0.07)
DIV	214 (13.0%)	1436 (87.0%)	1650	0.61 (±0.11)
MERGED	343 (18.3%)	1534 (81.7%)	1877	0.61 (±0.11)
EXT	132 (9.2%)	1304 (90.8%)	1436	0.57 (±0.12)

[a] # sub: number of substrates. [b] # nonsub: number of nonsubstrates. [c] # total: number of total compounds. [d] Average pairwise Tanimoto similarity based on MDLPublicKeys (±standard deviation)

Chemical Descriptors

P-glycoprotein is characterized by a promiscous ligand recognition. Global models thus most probably need to rely on rather general descriptors, such as lipophilicity, polar surface area, and number of H-bond donors and acceptors rather than on very specific ones. We thus used 11 simple physicochemical properties and pharmacophoric feature counts as descriptors for the development of our models. Furthermore, these descriptors encode molecular concepts that allow direct implementation in drug discovery and lead optimization projects. The molecular descriptors have been calculated using MOE2007.09.[21] A list of those descriptors along with a short description is given in Table 2. The chemical space spanned by these descriptors for our data sets is graphically represented in Figure 1. Symbols denote the different data sets. Figure 1 resembles the first two axes of a principal component analysis (PCA) on the three data sets, HOM, DIV, and EXT, explaining 76% of the variance in the data. It can be seen that all three data sets are overlapping in their chemical space and that, as expected, the HOM set (yellow squares) only populates part of the space of the other two sets.

Table 2: Descriptors used in this study.

Name	Description
a_acc	Number of hydrogen bond acceptor atoms
a_don	Number of hydrogen bond donor atoms
b_rotN	Number of rotatable bonds. A bond is rotatable if it has order 1, is not in a ring, and has at least two heavy neighbors.
logP(o/w)	Log of the octanol/water partition coefficient. This property is calculated from a linear atom type model with r2=0.931
mr	Molecular refractivity (including implicit hydrogens). This property is calculated from an 11 descriptor linear model with r2=0.997
PEOE_VSA_HYD	Total hydrophobic van der Waals surface area.
TPSA	Polar surface area calculated using group contributions to approximate the polar surface area from connection table
vsa_acc	Approximation to the sum of VDW surface areas (Å) of pure hydrogen bond acceptors
vsa_don	Approximation to the sum of VDW surface areas (Å) of pure hydrogen bond donor
vsa_hyd	Approximation to the sum of VDW surface areas (Å) of hydrophobic

	atoms.
Weight	Molecular weight (including implicit hydrogens) in atomic mass units

Classification Algorithm

In this study we apply the rule-based classification algorithm RuleFit, that has been recently presented by Friedman and Popescu.[20, 22] The algorithm is stated to combine high accuracy with good interpretability of the models obtained. It has recently been applied by the group of Jain to model chemical carcinogenicity and showed excellent performance in comparison to other machine learning algorithms like support vector machines and lazy learning methods.[23] Additionally, the

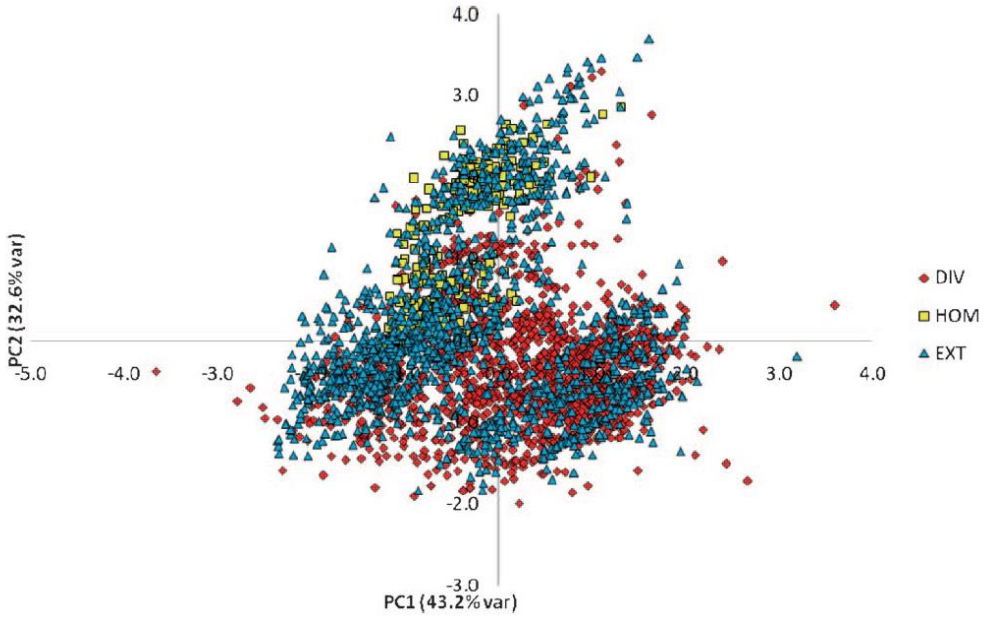


Figure 1: Chemical space oft he data sets used visualized by a PC score plot. HOM = yellow squares, DIV = red squares, EXT = blue triangles.

ability to interpret the results from the trained models was superior to most other methods. The RuleFit method belongs to the class of ensemble learning algorithms that establish classifiers on a usually low-bias, but high-variance ensemble of simpler classifiers. In the following section we give a brief overview on this method, for a more detailed description see the technical report by Friedman. Ensemble learning algorithms usually take the structural form of a linear model as expressed in Equation 1.

$$F(x) = a_0 + \sum_{m=1}^{M} a_m f_m(x)$$

where M is the size of the derived ensemble and $f_m(x)$ is a function of the training data constructed by each base learner (ensemble member). The ensemble predictions are usually made by linear combinations of the predictions of each of the individual base learners, with $\{a_m\}$ being the parameters specifying the linear combinations. The RuleFit method utilizes the importance sampled learning ensemble (ISLE) methodology for deriving the final ensemble

model. The ISLE methodology has been explained in Friedman and Popescu.[22] Ensembles of the type as shown in Equation 1 are built in RuleFit by using simple rules as base learners. The approach used for generating such rules is to view a decision tree ensemble as a collection of rules. That is, decision trees are used as base learners and each interior and terminal node of each resulting tree produces a conjunctive rule of the following form (Equation 2),

$$r_m(x) = \prod_{j=i}^{n} I(x_j \in s_{jm})$$

where n is the number of features in the rule, s_j represents the set of all possible values for the input feature vector x_i and s_{jm} is a specified subset of those values. Out of interpretation purposes a rule ensemble should be defined by simple rules and each rule should ideally be constructed by a small number of variables. Classification in RuleFit is done via two steps. In the first step an ensemble of decision trees is generated and consecutively the rules *r* are inferred from these decision trees. In the next step coefficients are fitted to these rules, which yields the parameters defined in Equation 3 shown below,

$$F(x) = a_0 + \sum_{k=1}^{K} a_k r_k(x)$$

where $\{a_k\}_0^K$ are the parameters that are determined during the training process. The details of this method are described in Friedman and Popescu.[20] During the training of the model the rules are pruned from the model, which implies that most of the coefficients $\{a_k\}_1^K$ from Equation 3 are set to zero. The magnitude of the remaining nonzero coefficients yields the relative importance of each rule. The sign of the coefficients determines whether a particular rule is associated with the positive (active) or the negative (inactive) class, in our case associated with ABCB1-substrate properties or ABCB1-nonsubstrate properties. The importance I_k of a rule is given in Equation 4.

$$I_k = |a_k| \times \sqrt{s_k(1 - s_k)}$$

The term S_k is described by Friedman as the so-called rule support, which reports the number of training examples in which the *k*-th rule holds. The support is defined in Equation 5.

$$s_k = \frac{1}{N}\sum_{i=1}^{N} r_k(x_i)$$

The importance of each particular input variable can be calculated by summing up the importances of the rules in which the particular input variable occurs (Equation 6).

$$J_i = \sum_{x_j \in r_k} I_k / m_k$$

The term m_k is introduced as a normalization weight to treat rules established on a different number of input variables equally. A feature of the RuleFit algorithm is that it only offers one parameter for tuning. This is the "path.inc" parameter, which controls the step size in the gradient search for model parameters. It has been shown by Popescu and also Jain that tuning

of this parameter does not improve model performance. Therefore, we kept this parameter at its default value of 0.2.

Model Validation and Performance Assessment

Performance of classification models was first internally evaluated by using 10-fold cross validation (10xCV) as well as 100 repetitions of Y-randomization (Yrand). For a description of these standard evaluation procedures see Tropsha et al.[25] Both evaluation procedures were carried out using self-written R scripts (the code is provided in the Supplementary material alongside with this article). In order to assess and compare the predictive power of the different models, several statistical parameters are reported. Five standard evaluation measures have been calculated from the obtained contingency tables (which are constituted by: TP=true positive; TN=true negative; FP=false positive; FN=false negative) to compare the different models. Since RuleFit returns a probabilistic prediction for each compound, we select compounds with a predicted probability higher then or equal to 0.5 to be classified as actives, whereas values smaller than 0.5 classify the compounds as inactives. The (i) correct classification rate (ccr) was calculated according to Equation 7:

$$ccr = 0.5\ x\ (\frac{TP}{TP + FN} + \frac{TN}{TN + FP})$$

Since we also have to deal with data sets that show a very unbalanced class distribution we prefer the formula given above for the overall correct classification rate, which also weights the class proportions, to the standard calculation (TP+TN)/(TP+TN+FP+FN), which does not consider uneven class distributions in the data set.[18] Furthermore, we apply Matthews correlation coefficient (mcc), which has emerged as a standard index for evaluating two-class machine learning models. It can be regarded as a discrete version of Pearson's correlation coefficient and can also be used even if the classes are of different sizes. The returned value can range from +1 to -1, where a coefficient of +1 indicates a perfect classification, a value of 0 indicates a random prediction and a coefficient of -1 indicates the worst possible prediction. It is calculated according to Equation 8:

$$mcc = \frac{(TP\ x\ TN) - (FP\ x\ FN)}{\sqrt{(TP + FP)\ x\ (TP + FN)\ x\ (TN + FP)\ x\ (TN + FN)}}$$

Additionally, we calculated (ii) sensitivity: TP/(TP+FN), (iii) specificity: TN/(TN+FP), (iv) precision on actives: TP/(TP+FP), and (v) precision on inactives: TN/(TN+FN). To give a graphical impression of the classification results also certain visualisation techniques, such as a Receiver Operating Characteristic (ROC), can be used.[26,27] On the x-axis a ROC plot displays the term 1 – specificity which corresponds to the "noise" in the data set, and on the y-axis it shows the sensitivity which can be interpreted as the "signal" that is to be identified by the ranking procedure. Contrary to that, an enrichment plot displays the amount of entries of the validation set on the x-axis and on the yaxis it shows the amount of instances from a given class retrieved from this validation set. One of the main differences between ROC plots and an enrichment plots is that enrichment plots only monitor the retrieval of actives of a screening experiment, whereas ROC plots also consider the ability of a model to discard inactives. Furthermore, the shape of the curve of the theoretical maximum in an enrichment

plot heavily depends on the ratio of actives to inactives in the data set, which is not the case for the ROC curve. Additionally, the area under the curve (AUC) spanned by the ROC curve can serve as a direct measure for comparing different models. Thus we select ROC plots as visualization tool for monitoring and comparing the classification results in this study. All our calculations presented in this article were carried out using the R implementation of the RuleFit method provided by Jerome H. Friedman (the executable binary has been obtained from: http://www-stat.stanford.edu/ ~jhf/R-RuleFit.html). The R/RuleFit version 8/10/05 with R 2.7.1 was used. The calculation of performance measures (accuracy, sensitivity, specificity,…) was done using self-written R functions. For generating ROC plots we used the ROCR package.[28]

Results and Discussion

Comparison of the Classification Performance of the Training Sets in Terms of Internal Predictivity

Table 3 displays commonly used standard measure of model performance with respect to internal validation. All three models show a relatively high ccr ranging from 0.75 for the DIV data set to 0.83 for the HOM data set. When assessing the class specific parameters, it can be seen that the HOM data set showed a balanced correct classification for the two classes. The sensitivity and the precision on actives for this set is 0.87 and 0.87. The specificity and the precision on inactives is 0.80 and 0.82, respectively. Contrary to that, the two other data sets, DIV and MERGED, showed a better classification for the inactive class, which is expressed as specificity and precision on inactives. For both data sets these values are >0.9. The sensitivity for these data sets is relatively low ranging from 0.54 for the DIV data set to 0.66 for the MERGED set. This most probably is due to the unbalanced ratio of substrates/nonsubstrates in the data sets. However, for these two sets the precision on actives is again satisfying (>0.7). Results of the Y-randomisation test are given in the last column of Table 3. These results show that none of the models gives a Y-randomized ccr higher than 0.63. When summarizing the results of this internal validation in terms of 10-fold CV and Y-randomisation it can be concluded that all three data sets gave reliable models with moderate Y-randomisation performance on the one hand, but with satisfactory performance in 10-fold CV. The full tables for each CV run as well as the contingency tables for each data set are given in Supplementary Information S1.

Table 3: Summary of classification results from internal validation (10-fold CV and Y-randomisation).

Set	Conf. mat.[a]				10-fold CV[b]						Yrand[c]
	TP[d]	TN[e]	FP[f]	FN[g]	ccr[h]	mcc[i]	sens[j]	spec[k]	pract[l]	prinact[m]	ccr
HOM	11	8	2	2	0.83	0.67	0.87	0.80	0.87	0.82	0.63
DIV	12	139	5	10	0.75	0.56	0.54	0.96	0.70	0.93	0.57
MERGED	23	145	8	11	0.81	0.64	0.66	0.95	0.73	0.93	0.54

[a] conf. mat.: confusion matrix. [b] CV: cross-validation. [c] Yrand: Y-randomisation. [d] TP: average number of true positives. [e] TN: average number of true negatives. [f] FP: average number of false positives. [g] FN: average number of false negatives. [h] ccr: correct classification rate. [i] mcc: Matthews correlation coefficient. [j] sens: sensitivity. [k] spec: specificity. [l] pract: precision on actives. [m] prinact: precision on inactives

Comparison of the Model Performance in Terms of External Predictivity

Additionally, we use our external validation set consisting of 1436 compounds to evaluate our three models. The resulting statistical parameters are reported in Table 4. The model derived from the MERGED data set performed best (ccr=0.9), whereas the model derived from a structurally homologous set of compounds (HOM), which gave the best results in internal validation, showed only a poor performance (ccr=0.55). This seems due to the fact that the HOM data set does not fully cover the chemical space of the external test set (see also Figure 1), and therefore most of the test compounds can be considered as being out of the applicability domain of this model. When assessing the NULL error rate for ccr for this external set, i.e. the ccr that would result if all compounds would be assigned to the larger (i.e. nonsubstrate class), a NULL-ccr of 0.5 would be derived. These results further underline the necessity of using external test sets for model evaluation. Further of interest is the fact, that the MERGED model showed only a poor classification on the actives in 10-fold CV, whereas these parameters are quite high when predicting the external set. This fact can be interpreted as the result of combining the HOM and the DIV data set to fully incorporate all available information on the ABCB1 substrates into the model. The DIV model showed an average performance with a ccr of 0.86 but only a weak precision on actives (0.45). Figure 2 shows the ROC plot for the three models when predicting the external validation set. The red diagonal ranging from the coordinates (0,0) to (1,1) corresponds to a random classifier, which is not able to distinguish signal from noise. Consequently, for any possible threshold the same percentage for sensitivity (signal) and 1-specificity (noise) is achieved. The AUC of the diagonal is 0.5 and any classifier above that limit can be considered as being better than random.

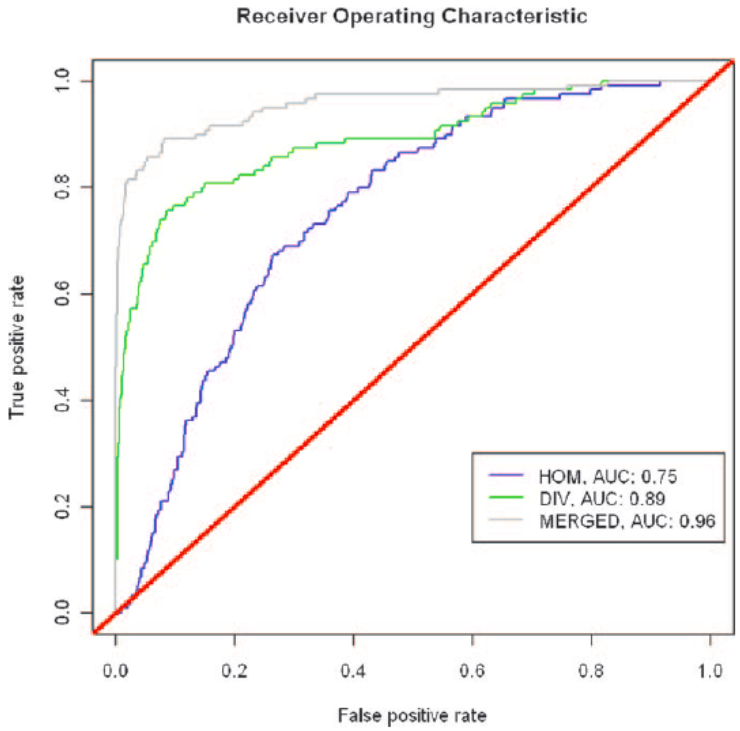


Figure 2: External validation: ROC plot

The ROC curve of an ideal classifier harbours an AUC of 1 and coincides at the coordinates (0,1), i.e. at the top left edge of the plot. From Figure 2 it can be seen that the three models behave quite differently. In a virtual screening situation the left side of the plot is most interesting, since it denotes how much signal (ABCB1-substrates) can be identified by the model while still discarding most of the noise. In the case of the MERGED model almost 80% of ABCB1 substrates can be retrieved without extracting false positives. The MERGED model also shows an AUC of 0.96. Again, as already indicated from the statistical evaluation (Table 4), the HOM model showed the poorest performance (AUC: 0.75), while the DIV model gives an average ROC curve (AUC: 0.89) compared to the other two models.

Table 4: Statistical evaluation of the classification for the external validation set

Model	TP[a]	TN[b]	FP[c]	FN[d]	ccr[e]	mcc[f]	sens[g]	spec[h]	pract[i]	prinact[j]
HOM	15	1293	11	117	0.55	0.23	0.11	0.99	0.97	0.25
DIV	102	1245	59	30	0.86	0.67	0.77	0.95	0.45	0.99
MERGED	107	1273	31	25	0.90	0.77	0.81	0.98	0.80	0.98

[a] TP: number of true positives. [b] TN: number of true negatives. [c] FP: number of false positives. [d] FN: number of false negatives. [e]ccr: correct classification rate. [f] mcc: Matthews correlation coefficient. [g] sens: sensitivity. [h] spec: specificity. [i] pract: precision on actives. [j] prinact: precision on inactives

Comparison to previously Published Models

Several authors have previously established in silico models to identify ABCB1 substrates. However, the fact that different data sets and different data set sizes as well as different descriptor types have been used in those studies makes a direct comparison almost

impossible. An overview of those machine learning studies which used a relatively large data set, is given in Table 5. Considering the data set size as well as the numbers of descriptors used, the model based on the MERGED data set compares favourable with the other studies.

Table 5: Overview of several in-silico models presented in prevous publications

Study	Data set			Test set performance	
	# comp	# desc	ML method	sens	spec
Gombar et al.[10]	140	27	LDA	0.94	0.78
Xue et al.[16]	201	22	SVM	0.84	0.67
De Cerquiera Lima et al.[18]	195	n.A.	CombiQSAR	0.78	0.84
Cabrera et al.[19]	203	5	TOPS-MODE	0.82	0.72
Huang et al.[12]	203	7	SVM	0.91	0.89
this study (HOM)	227	11	RuleFit	0.11	0.99
this study (DIV)	1650	11	RuleFit	0.77	0.95
this study (MERGED)	1877	11	RuleFit	0.81	0.98

Interpretation of the Models

Since the MERGED model consists of the most comprehensive collection of ABCB1 ligands available and it also gave the best performance in predicting the external data set, we restrict our analysis and interpretation to this model. Furthermore, a ranking of the obtained descriptor importances from RuleFit showed that most descriptors were equally ranked independent of the data set used to establish the models (Table 6). From this ranking of descriptor importances it can be seen that vsa_acc (100% relative importance, see Figure 3), which represent the sum of the van der Waals surface areas of all H-bond acceptors in a given molecule is the most important descriptor. The importance of H-bond acceptors in substrates and modulators of ABCB1 has also been highlighted by other studies.[9, 11, 29] The second most important descriptor for the MERGED model is logP(o/w), which shows a relative importance of 74.7% for the merged model. Since binding of ABCB1 substrates is supposed to take place at the membrane-embedded TMDs, the importance of lipophilicity for recognition and binding of xenobiotics has been experimentally shown in biochemical studies.[30] Raub et al. suggested that a reduction of lipophilicity might be a promising way to turn ABCB1 substrates into nonsubstrates.[31] Furthermore, numerous QSAR-studies on ABCB1 ligands also pinpoint the importance of logP as physicochemical descriptors.[32]

Table 6: Ranking of descriptor importances for the three models.

Desc	Model		
	HOM	**DIV**	**MERGED**
a_acc	*9*	*10*	*11*
a_don	6	*11*	*10*
b_rotN	7	8	5
logP(o/w)	**3**	5	**2**
mr	5	7	*9*
PEOE_VSA_HYD	4	4	**3**
TPSA	**2**	**2**	4
vsa_acc	**1**	**1**	**1**
vsa_don	*11*	*9*	6
vsa_hyd	*10*	6	7
Weight	8	**3**	8

The third most important descriptor in this model is PEOE_VSA_HYD (relative importance: 68.8 %, see Figure 3), which represents the total hydrophobic van der Waals surface area. It is noteworthy to mention that also TPSA, b_rotN, vsa_don, as well as vsa_hyd contain more than 50% in relative variable importance (Figure 3), which underlines that ABCB1 substrate recognition and binding does not appear to consist of one factor alone. The least important descriptors are the mere counts of H-bond acceptors and donors (a_acc and a_don). It is particularly interesting that the vsa_acc descriptor seems to be more predictive than the simple count of N's and O's. However, the inter-correlation of vsa_acc and a_acc as well as the inter-correlation of vsa_don with a_don is quite high (vsa_acc vs. a_acc: 0.78 and vsa_don vs a_don: 0.84), which might explain the low importance of descriptors that encode acceptor and donor properties, when they are represented as mere counts rather than as VSA projections, in the relative RuleFit variable ranking (Figure 3).

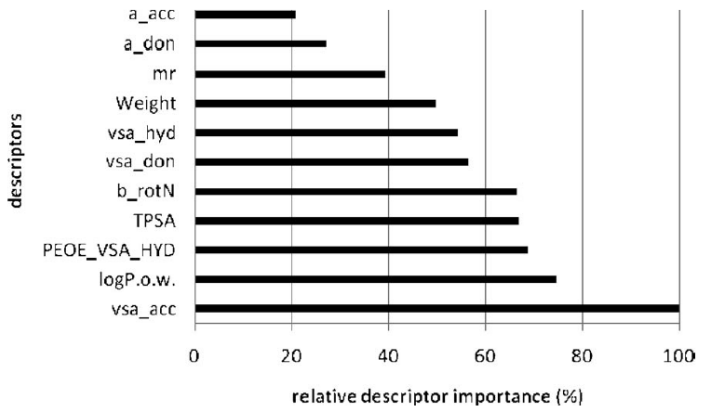


Figure 3: Relative descriptor importances for the MERGED model

Furthermore, the variable importances produced by a machine learning model can also be analyzed and verified by examining the correlation coefficients of the input variables with the class label (Figure 4). Although the mere correlations can only provide a limited understanding of the linear associations between molecular descriptors and the respective class label, the top ranked descriptors of our model showed also the highest correlations with the class label.

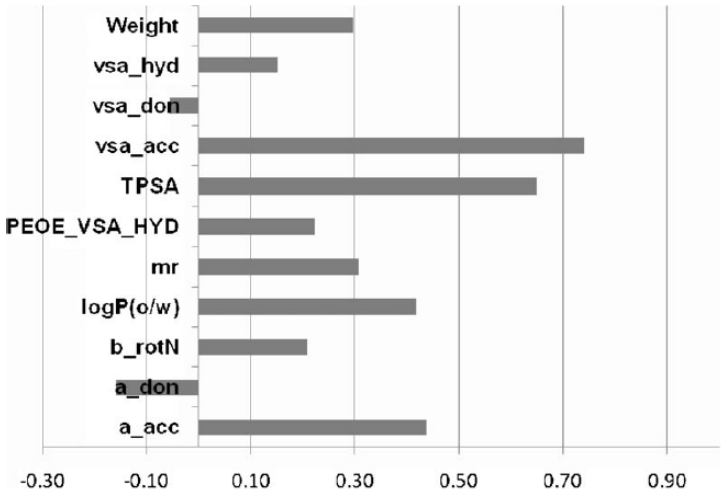


Figure 4: Pearsons vorrelation coefficients of the 11 descriptors with the class labelof the MERGED model.

Table 7 presents the five most important rules for the model, shown in order of their respective importance. The first column indicates if the rule characterizes the substrate class (S) or the nonsubstrate class (NS). The second column shows the rule importance scaled in that way that the highest estimate receives a value of 100. The last column resembles the respective rules. Note, that the numbers in squared brackets behind the descriptor denote the range of values this descriptor takes in our data set. Examining these most important rules reveals that nonsubstrates (rules 1 and 2) are characterized by low vsa_acc values (roughly <35), logP(o/w) values smaller than 4.8 and at least 2 H-bond donors. Contrary to that substrates show an H-bond acceptor van der Waals surface area higher than 37 (rule 4), are considered to be highly flexible (number of rotatable bonds >8), but show at maximum 4 H-bond donors. Furthermore, substrates have higher hydrophobic van der Waals surface areas

(PEOE_VSA_HYD) than nonsubstrates (rules 3 and 5) and seem to be bigger in size and shape (rule 5; mr>app.14).

Table 7: Five most important rules produced by the MERGED model. Numbers in [] denote the respective data range

Class	Imp.	Rule
NS	100	vsa_acc [2.5;77.5] ≤ 34.71 AND logP(o/w) [-1;7.7] ≤ 4.79
NS	92.9	a_don [0;9] ≥ 2 AND vsa_acc [2.5;77.5] ≤ 34.71
S	64.8	PEOE_VSA_HYD [89;668] ≥ 436.1 AND 102 ≤ TPSA [44;185]_127.3
S	47.0	b_rotN [1;9] ≥ 8 AND a_don [0;9] ≤ 4 AND vsa_acc [2.5;77.5] ≥ 36.81
S	35.1	505.6 ≤ Weight [155;728] ≥ 550.3 AND PEOE_VSA_HYD [89;668] > 403.8 AND mr [4;19] ≥13.9

Conclusion

In drug discovery and development, in silico approaches have the potential to provide cost-effective screening tools for identification of compounds with unfavourable ADMET properties. However, some of these in silico approaches either utilize "black-box" algorithms or make use of highly complex descriptors, which renders interpretation and also the translation of retrieved patterns of the models a highly difficult and maybe also misleading task. We utilize RuleFit modelling as a computational method for in silico screening in order to classify compounds with respect to their ABCB1 substrate properties. The derived models are based on simple, physicochemical descriptors, which are on the one hand fast and easy to calculate and on the other hand encode simple chemical concepts that are familiar to the medicinal chemist. The good performance of the models obtained highlights the applicability of RuleFit modelling for classification of ABCB1 substrates and thus for identification of potential ADMET risks. It is noteworthy to mention that the best results were achieved with the structurally diverse MERGED data set. This should allow a permanent dynamic updating of the model with new chemical scaffolds from new lead optimization series. The algorithmic nature of the RuleFit method in conjunction with a set of simple physicochemical properties allowed to derive basic rules that can assist in guiding lead optimization projects. In summary, the most important rules derived suggest that substrates are associated with large hydrophobic surfaces, expressed as PEOE_VSA_HYD, are quite flexible and show less than four H-bond donors. Contrary to that, nonsubstrates are mainly characterized by small vsa_acc and logP(o/w) values and are enriched in Hbond donors. The presented method allowed accurate and fast classification of ABCB1 substrates for a highly diverse set of hit and lead compounds from the Boehringer Ingelheim inhouse library. RuleFit modelling also proved to be very useful in simplifying the interpretation of the models and helped in identification of molecular properties that characterize potential ABCB1 substrates and nonsubstrates.

Acknowledgements

We gratefully acknowledge financial support by the Austrian Research Promotion Agency (FFG); grant: #BI-812074) and by the Austrian Science Fund (SFB # 3502).

References

[1] R. Pearlstein, R. J. Vaz, D. Rampe, J. Med. Chem. 2003, 46, 2017–22.
[2] M. M. Gottesman, T. Fojo, S. E. Bates, Nat. Rev. Drug Disc. 2002, 2, 48 –58.
[3] G. Szak_cs, A. Varadi, C. Ozvegy-Laczka, B. Sarkadi, Drug Disc. Today 2008, 13, 379–393.
[4] C. F. Higgins, K. J. Linton, Nat. Struct. Mol. Biol. , 2004, 11, 918 – 926.
[5] R. J. Vaz, T. Klabunde, Methods and Principles in Medicinal Chemistry, Vol. 38, Antitargets – Prediction and Prevention of Drug Side Effects (Ed: P. D. R. Mannhold, H. Kubinyi, G. Folkers), Wiley-VCH, Weinheim, 2008.
[6] M. A. Demel, R. Schwaha, O. Kr_mer, P. Ettmayer, E. E. Haaksma, G. F. Ecker, Expert Opin. Drug. Metab. Toxicol. 2008, 4,1167–1180.
[7] A. Seelig, A. , Eur. J. Biochem. 1998, 251, 252–261.
[8] A. Seelig, E. Landwojtowicz, Eur. J. Pharm. Sci. 2000, 12, 31–40.
[9] R. Didziapetris, P. Japertas, A. Avdeef, A. Petrauskas, J. Drug. Target. 2003, 11, 391–406.
[10] V. K. Gombar, J. W. Poli, J. E. Humphreys, S. A. Wring, C. S. Serabjit-Singh, J. Pharm. Sci. 2004, 93, 957–968.
[11] M. P. Gleeson, J. Med. Chem. 2008, 28, 817–34.
[12] J. Huang, G. Ma, I. Muhammed, Y. Cheng, J. Chem. Inf. Model. 2007, 47, 1638–47.
[13] M. A. Cabrera, I. Gonzalez C. Fernandez, C. Navarro, M. Bermejo, J. Pharm. Sci. 2006, 95, 589–606.
[14] J. Cartmell, S. Enoch, D. Krstajic, D. E. Jeahy J. Comput. Aided Mol. Des. 2005, 19, 821–833.
[15] Y. Xue, J. Chem. Inf. Model. 2004, 44, 1630–1638.
[16] Y. Xue, J. Chem. Inf. Model. 2004, 44, 1497–1505.
[17] R. Wang, J. Clin. Pharm. Therapeutics 2003, 28, 203–228.
[18] P. De Cerqueira-Lima, J. Chem. Inf. Model. 2006, 46, 1245–1254.
[19] M. A. Cabrera, J. Pharm. Sci. 2005, 95, 589–606.
[20] J. H. Friedman, B. E. Popescu, Predictive Learning via Rule Ensembles, Technical Report, 2005.
[21] MOE2007.09, C.C.G.I., http://www.chemcomp.com.
[22] J. Sadowski, H. Kubinyi, J. Med. Chem., 1998, 41, 3325–3329.
[23] J. H. Friedman, B. E. Popescu, Gradient Directed Regularization
for Linear Regression and Classification, Technical Report, 2004.
[24] J. J. Langham, A. N. Jain, J. Chem. Inf. Model. 2008, 48, 1833–1839.
[25] K Baumann, Quant. Struct.-Act. Relat. 2002, 21, 507–519.
[26] J. P. Egan, Signal Detection Theory and ROC Analysis, Series in Cognitition and Perception, Academic Press, New York, 1975.
[27] A. Hillebrecht, G. Klebe, J. Chem. Inf. Model. 2008, 48, 384–396.
[28] T. Sing, Bioinformatics 2005, 21, 3940–3941.
[29] P. Chiba, G. Ecker, D. Schmid, J. Drach, B. Tell, S. Goldenberg, V. Gekeler, Mol. Pharmacol. 1996, 49, 1122–1130.
[30] A. Rothnie, J. Storm, R. McMahon, A. Taylor, I. D. Kerr, R. Callaghan , J. Biol. Chem. 2005, 279, 3984–90.
[31] T. J. Raub , Mol. Pharmaceutics 2006,3, 3–25.

[32] K. Pleban, G. F. Ecker, Minirev. Med. Chem. 2005, 5, 153–165.

Author Contribution

This chapter was published in *Molecular Informatics* as

Demel M., Krämer O., Ettmayer P., Haaksma E., Ecker G.F.**Ensemble Rule-Based Classification of Substrates of the Human ABC-Transporter ABCB1 Using Simple Physicochemical Descriptors.** *Molecular Informatics*, 2010, 29, 233-242.

The BIA sets were provided by Boehringer Ingelheim Austria (Dr. Oliver Krämer). Dr. Krämer also provided the external test set. Dr. Haaksma, Dr. Ettmayer and Dr. Krämer assisted in the design of computational experiments. I performed all the calculations. The results were evaluated by me and Dr. Krämer. The initial manuscript for publication was written by me and reviewed, edited and improved by all the co-authors. Prof. Ecker supervised this project.

Supplementary Information

Supplementary Information S1:

Internal Evaluation of the models by 10-fold CV

Table S1.1 10-fold CV classification results for the HOM data set

CV run	tp	tn	fp	fn	acc	sens	spec	pract	prinact	ccr
1	11	10	1	0	0.95	1.00	0.91	0.92	1.00	0.95
2	11	6	4	1	0.77	0.92	0.60	0.73	0.86	0.76
3	8	8	1	5	0.73	0.62	0.89	0.89	0.62	0.75
4	10	7	3	2	0.77	0.83	0.70	0.77	0.78	0.77
5	12	6	3	1	0.82	0.92	0.67	0.80	0.86	0.79
6	11	8	2	1	0.86	0.92	0.80	0.85	0.89	0.86
7	8	12	0	2	0.91	0.80	1.00	1.00	0.86	0.90
8	13	8	1	0	0.95	1.00	0.89	0.93	1.00	0.94
9	13	4	1	4	0.77	0.76	0.80	0.93	0.50	0.78
10	12	6	2	1	0.86	0.92	0.75	0.86	0.86	0.84
mean	11	8	2	2	0.84	0.87	0.80	0.87	0.82	0.83
SD	2	2	1	2	0.08	0.12	0.12	0.08	0.16	0.08

tp=true positives, tn=true neagatives, fp=false positives, fn=false negatives, acc=overall accuracy ((TP+TN)/(TP+TN+FP+FN)), sens=sensitivity, spec=specificity, pract=precision on actives, precision on inactives, ccr=correct classification rate.

Table S1.2 10-fold CV classification results for the DIV data set

CV run	tp	tn	fp	fn	acc	sens	spec	pract	prinact	ccr

1	12	143	4	7	0.93	0.63	0.97	0.75	0.95	0.80
2	16	134	8	8	0.90	0.67	0.94	0.67	0.94	0.81
3	10	145	4	7	0.93	0.59	0.97	0.71	0.95	0.78
4	12	133	7	14	0.87	0.46	0.95	0.63	0.90	0.71
5	17	137	5	6	0.93	0.74	0.96	0.77	0.96	0.85
6	13	137	2	13	0.91	0.50	0.99	0.87	0.91	0.74
7	11	140	2	12	0.92	0.48	0.99	0.85	0.92	0.73
8	6	143	8	8	0.90	0.43	0.95	0.43	0.95	0.69
9	10	137	3	15	0.89	0.40	0.98	0.77	0.90	0.69
10	9	138	8	10	0.89	0.47	0.95	0.53	0.93	0.71
mean	12	139	5	10	0.91	0.54	0.96	0.70	0.93	0.75
SD	3	4	2	3	0.02	0.11	0.02	0.14	0.02	0.06

tp=true positives, tn=true neagatives, fp=false positives, fn=false negatives, acc=overall accuracy ((TP+TN)/(TP+TN+FP+FN)), sens=sensitivity, spec=specificity, pract=precision on actives, precision on inactives, ccr=correct classification rate.

Table S1.3 10-fold CV classification results for the MERGED data set

CV run	tp	tn	fp	fn	acc	sens	spec	pract	prinact	ccr
1	20	145	7	16	0.88	0.56	0.95	0.74	0.90	0.75
2	25	145	6	12	0.90	0.68	0.96	0.81	0.92	0.82
3	21	150	6	11	0.91	0.66	0.96	0.78	0.93	0.81
4	27	142	6	12	0.90	0.69	0.96	0.82	0.92	0.83
5	17	147	10	13	0.88	0.57	0.94	0.63	0.92	0.75
6	22	144	10	11	0.89	0.67	0.94	0.69	0.93	0.80
7	24	144	8	11	0.90	0.69	0.95	0.75	0.93	0.82
8	31	142	6	8	0.93	0.79	0.96	0.84	0.95	0.88
9	23	138	15	11	0.86	0.68	0.90	0.61	0.93	0.79
10	18	152	8	9	0.91	0.67	0.95	0.69	0.94	0.81
mean	23	145	8	11	0.90	0.66	0.95	0.73	0.93	0.81
SD	4	4	3	2	0.02	0.07	0.02	0.08	0.01	0.04

tp=true positives, tn=true neagatives, fp=false positives, fn=false negatives, acc=overall accuracy ((TP+TN)/(TP+TN+FP+FN)), sens=sensitivity, spec=specificity, pract=precision on actives, precision on inactives, ccr=correct classification rate.

Chapter 13: Manuscript I - Random Forests QSAR models for ABCB1 substrate/nonsubstrate Classification: Attempts to bridge the Efficacy/Effectiveness gap

Abstract

ABCB1 (P-glycoprotein, P-gp) represents the paradigm member of the human ATP-binding cassette (ABC)-transporter family which are integral membrane proteins that utilize the energy provided from ATP binding and hydrolysis to actively extrude drugs out of the interior of cells against a steep concentration gradient. ABCB1, as a multidrug efflux pump, that recognizes a wide variety of chemically and functionally unrelated compounds is on one hand able to confer a multidrug resistant (MDR) phenotype to cancer cells but on the other hand also represents a protective mechanism for the human body against xenobiotics and potential harmful substances. Thereby ABCB1 critically influences the pharmacokinetic fate of drugs that are substrates of it. Hence, its (over)expression is associated with reduced chemotherapeutic treatment success and also with limited bioavailability of drugs and with clinically relevant drug-drug interactions. *In-silico* classification models provide versatile means for the detection of ABCB1 substrates early in the drug development process. Nowadays, a variety of such classification models are present in the literature. Most of these models show satisfying performance and their *efficacy* has been evaluated using *a-priori* defined test sets. However, their *effectiveness* (=generalizability) has not been proven in many cases. It is the primary objective of this study to critically appraise the performance of different ABCB1 substrate/non-substrate classification models with respect to their capability to predict new molecules that origin from different sources. In order to accomplish this task we utilize data that are derived from three independent sources and described by three different descriptor sets. These data sets were subjected to Random Forest modeling. An external test set was used to determine the effectiveness of the established models. It is shown that for certain models excellent results were achieved. However, some models clearly fail to reliably predict this external test set, which unravels a potential "*efficacy-effectiveness gap"* of ABCB1 substrate/non-substrate classification models. To account for this, two different measures for assessing the applicability domain were applied. These post-processing techniques are shown to be useful for improving the prediction accuracy of weak models. Nevertheless, this gain in performance is accompanied with a large amount of molecules considered to be out of the model domain and therefore remain as being virtually unclassified.

Introduction

Human multidrug ATP-binding-cassette (ABC)-transporters comprise a family of membrane-bound efflux pumps, that utilize the energy of ATP binding and hydrolysis to export xenobiotics out of living cells [1]. The human genome encodes for 48 functional ABC-transporters that are categorized into seven subfamilies (ABC-A to ABC-G) on basis of their sequence similarity. ABCB1 (P-glycoprotein, Pgp) is the best characterized member of these multidrug efflux proteins and has been the center of many biochemical and pharmacological investigations in the last three decades [2]. ABCB1 when being overexpressed confers a multidrug resistant (MDR) phenotype to cancer cells by actively extruding cytostatic drugs against a steep concentration gradient out of cancer cells. Additionally, ABCB1 was also

found to be physiologically expressed in normal tissue. This expression pattern explains the physiological function of ABCB1 as representing a protective mechanism for the human body against xenobiotics and shows that ABCB1 is an essential contributor to the *a*dminstration-, the *d*istribution- and the *e*limination- phases of the prominent pharmacokinetic *administration-distribution-metabolisation-elimination (ADME)* paradigm [3], [4]. In fact, ABCB1-mediated drug efflux has been identified as the main reason for limited bioavailability for many drugs and as a crucial mechanism for several drug-drug interactions. Pharmacological studies have shown that this transporter is capable to bind and transport a wide variety of chemical and pharmacological unrelated compounds that are generally described as being large and hydrophobic [5].

The experimental transport activity assessment of ABCB1 substrates can be achieved through the application of a variety of *in-vitro* assays. However, these experimental assays are often used in late stages of the drug development process, when compounds are already optimized towards classical "drug-like" properties such as potency, solubility, or absorption [6]. Furthermore, these experimental procedures are rather expensive and time-consuming. Taken this together, the establishment of accurate, fast and cost-effective *in-silico* models represents an urgent need in contemporary MDR and ABCB1-related ADME profiling [7]. Due to the lack of a high-resolution x-ray structure of human ABCB1, ligand-based QSAR-classification models currently represent the current state-of-the-art techniques to identify potential ABCB1 substrates early in the drug design process [8].

Much work has been invested so far to provide classification systems regarding ABCB1-substrates and non-substrates. These efforts range from the development of a variety of pharmacophore models, and simple and intuitive "rule-based" approaches to highly sophisticated machine learning models. A comprehensive overview of these approaches is given in [8]. Most of the rule-based computational methods that have been developed to characterize ABCB1-substrates provide simple and direct guidelines for medicinal chemists to design-in or design-out substrate properties. However, most of them have never been validated on external test sets. Recently we utilized a rather large and chemically highly diverse collection of proprietary Boehringer Ingelheim in-house hit- and lead-like compounds to construct an ensemble rule-based classification model that achieved a correct classification rate of 90% on a defined test set consisting of 1436 compounds [9]. Additionally, in 2011, Wang and colleagues presented the so far most comprehensive collection of public available compounds annotated as ABCB1 substrates and non-substrates in the literature. They developed a RBF-kernel based SVM and used autocorrelation and MOE descriptors to construct a model that showed an overall accuracy of 88% on an external test set of 120 compounds.

As highlighted above a vast collection of ligand-based classification models are available, that in general show satisfying classification performance. It needs to be stressed out, that the main utility of such QSAR classification models is the prediction of new molecules. Thus, the model needs thorough validation in order to be successfully applied in real-life situations. Various validation procedures are mentioned in the literature. Roughly, these procedures can be divided into internal and external methods. Internal validation

techniques are usually based on iterative data set splitting, i.e. certain molecules of the data set are excluded from the model calculation. 10-fold crossvalidation is one of the most known and used methods for assessment of the internal validity of a classification model [10]. In contrast to internal validation, external validation techniques make use of new data that has not been used in the modeling process. Ideally, such new molecules shall be derived from external sources to provide ultimate information on the external validity, also referred to as generalizability, of the model. Unfortunately, for most modeling tasks data are scarce and therefore, such an external test set is usually defined *a-priori*, and validation is done by classifying these "supposedly unknown" data points, rather than utilizing data that origin from "real external" sources. Although, the *a-priori* definition of training and test set represent generally accepted means as surrogate validation criteria for *in-silico* models, many models have been shown to show decreased predictive performance when being applied in practical situations. We refer to this decrease in performance as the "*efficacy-effectiveness gap*". This terminology is adapted as an analogy from the field of clinical drug evaluation. In many cases it is observed that drugs do not perform as well in clinical practice as in clinical drug trials that have been the basis of their benefit-risk evaluation and their consecutive market authorization [11]. Thus, they show a decreased *effectiveness*, whereas clinical trials predicted good *efficacy*. This decrease in *effectiveness* of clinically used drugs often relies upon the fact that drug trials mostly reflect ideal circumstances in which only a rather small and often highly homogenuos collective of patients is investigated. In analogy to this clinical situation we encounter a similar problem in QSAR classification modeling. Models are mostly evaluated on small, homogeneous sets of molecules (i.e. they show a good *efficacy*), but might deliver disappointing results when applied to the whole drug-like universe (i.e. they show a poor *effectiveness*). In this context *efficacy* describes the performance of a model on a *a-priori* defined test set, whereas *effectiveness* refers to the performance of the model when being applied to data that origin from external sources. Most of the models that address the ABCB1 substrate/non-substrate classification problem are based on rather small data sets and have only been evaluated on *a-priori* defined test sets that mostly resemble only the study-specific, defined chemical space. The general applicability for the whole drug-like chemical space of these models has not been proven so far. In an attempt to address this question, it is the primary objective of this study to challenge ABCB1 substrate/non-substrate classification models that show a satisfying performance when evaluated on basis of internal validation techniques, with newly derived test sets that origin from different data sources. Furthermore, post-processing techniques that aim to estimate the applicability domain of the models are also employed and their feasibility is critically appraised.

Material and Methods

Data Sets.

In this study we utilize three different data sources that together are compiled to yield 9 training data sets and one test data set. All the sets contain binary dependent variables [1,0] that denote ABCB1-substrates [1] and ABCB1 non-substrates [0]. The first data set was previously used in Demel et al. [9] and comprises a highly diverse collection from different Boehringer Ingelheim Austria (BIA) hit-to-lead optimization projects. The ABCB1 substrate

liability was assessed using a cytotoxicity assay, that is based on the ratio of the EC50 value of a compound in a ABCB1 overexpressing cell line and its EC50 value in the wild-type cell line. According to this, ABCB1 substrates that display a decreased activity in the transporter overexpressing cell line, exhibit a higher ratio of the two compared EC50 values (for more details see Ref. [9]) The whole training data from the Boehringer Ingelheim Austria (BIA) library comprises 1877 compounds and an additional test set, which is comprised of 1436 compounds. The training data are derived from two smaller BIA-in house data sets. The first one resembles a small collection of 227 very homogenous compounds (BIAHOM), whereas the second set resembles a highly diverse data set consisting of 1650 compounds (BIADIV). The external set from BIA (EXT) consists of 1436 molecules. The BIA sets are primarily characterized by "classical man-made, synthetic" hit- and lead-like chemicals that are enriched in nitrogen containing ring scaffolds. Furthermore, the BIA sets are highly unbalanced with respect to the class distribution of their substrates and non-substrates. In an attempt to establish also a more balanced BIA set, BIAMD, which is constituted of all 343 BIA substrates and also of 343 randomly sampled inactive molecules was constructed. Noteworthy, the 343 inactive, non-substrates were randomly selected from the other inactive molecules of the BIA training library to also yield a more diverse, but balanced BIA set. The second data set is derived from the NCI60 ABC-transporter drug sensitivity screen in which mRNA levels of all human multidrug ABC-transporters were correlated with the growth inhibitory effect of a compound across 60 different cancer cell lines [12]. According to this approach a compound can be considered as a substrate if it shows a negative correlation, whereas compounds that exhibit a non-correlation can be considered as non-substrates. We utilize data for the ABCB1 transporter from this screen. This data set (NCI) consists in total of 240 compounds and was previously used by us in a study which aimed to address the feasibility of different data pre-processing techniques in the context of *k*NN-classification QSAR approaches [13]. Compounds with correlations lower than -0.3 were assigned to be substrates and compounds with correlations in the range of $-0.02<r<+0.02$ were assigned to be non-substrates. In total 240 molecules that mainly resemble natural products or derivatives of these were used from this NCI60 screen. The last data set consists of a comprehensive collection of compounds obtained from different literature sources to yield a final set of 257 molecules (LIT) [14–19]. The LIT set is primarily composed of marketed drugs (digoxin, loperamide, antiretroviral agents, cholesterol-lowering agents) or molecular probes commonly used in contemporary pharmacology, such as fluorescent dyes (e.g. Calcein-AM). Another striking difference between the other two data sets (BIA, NCI) is that most of the compounds in the LIT set have been pharmacologically evaluated in classical transporter assays (e.g. efflux assay, ATPase assay), rather than in a cytotoxicity assay. The merged set of the NCI and the LIT set constituted the LITNCI set consisting of 497 molecules. Taken together, all these compounds derived from these three sources embody a total of 2375 training molecules (global). A tabular summary of all these data sets is given below.

Table 1: Data sets derived from different sources used for RF modelling

Set	# substrates	# non-substrates	# total	characteristics

BIAHOM	129	98	227	drug-like hits and leads with inherent cytotoxicity
BIADIV	214	1436	1650	
BIADIVHOM	343	1534	1877	
BIAMD	343	343	686	
NCI	110	130	240	natural product derived cytotxics
LIT	145	112	257	marketed drugs assayed in transporter assays
LITNCI	255	242	497	sets merged from different sources
BIAMDLITNCI	598	585	1183	
Global	*599*	*1776*	*2375*	

Descriptors

In order to adequately cover different aspects of the contained chemical information of our data sets we utilize three different descriptor sets as numeric representations of our molecules. The first descriptor set contains a collection of 77 Autocorrelation (ACORR) descriptors. The second descriptor set is constituted of the 32 VSA descriptors (VSA) developed by Labute [20]. The third numeric representation consists of the calculation of 166 MACCS atom-type count descriptors (MACCS).

Random Forest Classification

In this study we apply Random Forests (RF) modeling to develop predictive classification models for our data sets. RF is a classification and regression algorithm introduced by Leo Breiman in 2000 [21]. RF belongs to the class of ensemble learning algorithms (similar to boosting and bagging) and as such represents a collection of *n* decision trees. It has been developed in order to overcome the poor predictive power of single decision trees, while retaining the appealing interpretation properties of the latter. The algorithm is explained in detail in Refs. [21], [22]. Briefly, RF initially performs a feature and instance reduction of the original training data set. It draws a bootstrap sample of the training data and selects randomly a pre-specified number (m_{try}) of variables. This reduced data matrix is used for generation of a CART decision tree that is fully grown without pruning. This procedure is repeated until *n* trees, that are of low bias, while at the same time are highly variant, are constructed. Additionally, RF also performs some type of internal validation parallel to training. This is done by applying the so called *out-of bag (oob)* method. Simply, the compounds that are not included by the bootstrap sample (the *oob* compounds) of the *n*-th tree are used to estimate the prediction performance of this tree. Finally, all votes for each *oob*-compound are collected in order to report the final training performance (average *oob* error)

for the ensemble. In general, ensemble machine learning models are usually considered to be "black box" models that are difficult to interpret. However, RF, as a collection of simple, easy-interpretable decision trees, offers tools to assess descriptor importance and thereby can also be used for model interpretation. RF descriptor importance is calculated as follows during training. Each tree is grown and predictions are made on the *oob* compounds for that tree. Simultaneously, each variable in the *oob* set is randomly permuted, one at a time, and each of these permuted instances are also predicted by the tree. If a variable is important for the model, it is expected that *oob* prediction accuracy decreases, when this variable is permuted. Contrary, a variable that is of minor importance to the model might retain a comparable prediction accuracy, when used with original values and when used with permuted values. Finally, all variables are ranked according to their differences in prediction accuracy when being permuted and when used with original values. The descriptor with the highest difference can be considered as the most important one. Owing to this appealing features of the RF algorithm, it has been used widely in the pharmacoinformatic literature [22–25].

Performance Assessment

The performance of the RF classification models produced in this study was assessed on basis of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). For RF models, compounds with a predicted positive class probability >0.5 have been assigned to the non-substrate class, whereas compounds with a predicted positive class probability <= 0.5 have been assigned to the substrate class. From the resulting confusion matrix, which contains the mere counts of the predicted classes, the following statistical parameters were calculated to assess model quality:

(i) *Matthews correlation coefficient (MCC)* which represents a discrete variation of Pearson`s correlation coefficient and is considered a standard measure if classes are of different sizes. It is calculated according to the following equation.

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

The calculated value can range from +1 to -1, where a value of +1 indicates a perfect classification, a value of 0 indicates a random prediction and a coefficient of -1 indicates the worst possible prediction.

(ii) *sensitivity (SE)*, the proportion of correctly identified actives among all actives.

$$SE = \frac{TP}{(TP + FN)}$$

(iii) *specificity (SP)*, the proportion of correctly identified inactives among all inactives,

$$SP = \frac{TN}{(TN + FP)}$$

(iv) *precision on actives (PRACT)*, the proportion of correctly identified actives among all compounds predicted as active,

$$PRACT = \frac{TP}{(TP + FP)}$$

(v) *precision on inactives (PRINACT)*, the proportion of correctly identified inactives among all compounds predicted as inactive

$$PRINACT = \frac{TN}{(TN + FN)}$$

Furthermore, we used the inherent oob-error rate of RF for internal validation of our models as well as Receiver Operating Characteristic (ROC) curves for external validation to visualize model performance [26].

Distance-to-model measures as post-processing techniques

Two different types of distance-to-model (D2M) measures are calculated in order to perform post-processing of the classification models for the external test set. Both are model-independent measures that belong to the class of descriptor-based AD measures.

Range.tsh. The first D2M measure that is applied in this study is one of the simplest approaches to estimate the convex hull of the training compounds by taking ranges of the individual descriptors. These ranges are defined as the minimum and maximum descriptor values of the training instances for each descriptor [27]. A graphical depiction of this measure is depicted in the figure below. These ranges define a *n*-dimensional rectangle with sides parallel to the coordinate axes. A compound of the test set is considered to be out of space if it is outside this multidimensional rectangle. In order to account for a models potential to extrapolate out of the spanned descriptor space, we consider a test instance to be "out of domain" if its descriptor values exceed the defined boundaries in more than 10% of all descriptors. This measure is rather simple to compute but harbors the disadvantage that it may enclose "empty space" within the defined AD, if the input data are not uniformly distributed. In the following we refer to this measure as "*range.tsh*".

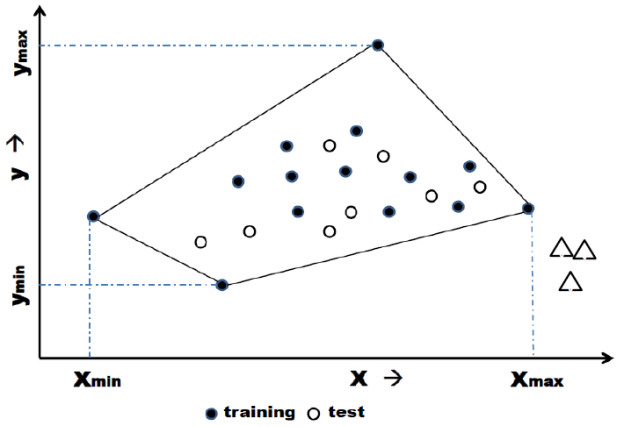


Figure 1: Theoretical considerations on the implementation of range.tsh, a descriptor-based AD measure, visualized on a simplified, hypothetical 2D example. Training instances, which are used to define the AD of the model are shown in black circles, whereas test instances inside the applicability domain are shown as white circles. Three test instances (white triangles) of the depiction are considered to be out of model space.

Dist.tsh. The second measure makes use of similiarity measurements to define the applicability domain of a model. It calculates the Euclidean distances among all training compounds and the test compounds. Then the distance of a given test instance to its nearest neighbour in the training set is compared to the defined applicability domain threshold. The classification result of a test instance is considered to be unreliable it the distance exceeds this threshold. This threshold is defined in the following way:

$$dist.tsh = y + Z\sigma$$

The parameters of this equation are defined as follows (see also [28–30]). At first, the Euclidean distances are compared in a pairwise fashion for the training instances and are then averaged. Then the set of all distances that is lower than this average is formulated. The two parameters, y and σ, are then determined as the average and the standard deviation of the distances of all compounds included in this set. The third parameter Z is an empirical cutoff value that is in this study chosen to be 0.5. A cutoff of $Z=0.5$ formally places the boundary for which compounds are predicted to be "out of domain" at one-half of the standard deviation (thereby a Boltzmann distribution of the distances between each compound and its neighbors is assumed) (29). In the following we refer to this measure as "*dist.tsh*".

Software

For the calculation of chemical descriptors, the Molecular Operating Environment (MOE) software (VSA and MACCS descriptors) was used. Additionally, for the calculation of autocorrelation vectors, the ADRIANA package from Molecular Networks was used. Modeling was done using the R-software environment. RF models were generated using the *randomForest()* package obtained from CRAN. ROC curves were calculated using the R-built-in functions of the *ROCR* package. Distance-to-model measures were calculated using self-written in-house functions in R language (see supplementary information).

Results and Discussion

Assessing the "efficacy" (internal validaty) of RF-ABCB1 classification models using oob error rate curves.

In this study all nine data sets described by the three descriptors types were subjected to Random Forest (RF) modeling. Each RF model was generated using 500 trees (*ntree*=500) and the parameter m_{try} which controls the number of descriptors that are used for the generation of each tree was kept at its default value ($m_{try} = \sqrt{n}, n = number\ of\ descriptors$). In total 27 Random Forest classification models were constructed. As described above, RF provides an inherent measure of internal validity that is calculated "on the fly" during model generation. This measure is termed the *oob*-error and reports the misclassification of each model. Hence, an oob-error rate of 0.1 (10%) corresponds to a correct overall classification accuracy of 0.9 (90%). In this study, we make direct use of this measure for determining the efficacy of the 27 models. A tabular overview of the average *oob*-error rate for all models per descriptor type is given in Table 2.

Table 2: Assessment of the training performance: Average oob-error rate for all models reported in % misclassification. The best models are shown in bold, whereas the worst performing models are highlighted in italics.

	ACORR	MACCS	VSA
bi_div	7.45%	5.88%	7.92%
bi_hom	15.42%	9.25%	8.81%
bidivbihom	8.58%	6.98%	8.08%
bimd	17.97%	**4.35%**	11.52%
bimdlitnci	*20.22%*	12.30%	12.64%
global	11.84%	10.36%	11.52%
lit	19.18%	14.51%	14.12%
litnci	15.15%	14.35%	12.74%
nci	10.42%	19.58%	19.17%

From this it can be seen that almost all model achieve a stable training classification error smaller than 20% (corresponding to < 0.2) after 500 trees. The only exception is the merged BIMDLITNCI set when described using ACORR descriptors which received an average *oob*-error rate of 20.22%. The model with the smallest *oob*-error is the BIMD model based on MACCS descriptors which exhibits an average *oob*-error of 4.35%. The global model, which contains the merged data sets and embodies more than 2300 molecules achieves *oob*-error rates of 11.84% for ACORR descriptors, of 11.52% for VSA descriptors and of 10.36% for MACCS descriptors. It is of interest to note that models solely based on BI data in general seem to perform better than the models that are constructed on the LIT or the NCI60 data set. A potential reason for that is maybe that this set does not suffer from inter-laboratory differences in assay measurements and that the cutoff for assigning ABCB1 substrates and ABCB1 non-substrates was set very rigorously. With respect to the three different descriptor types that were used, no preference can be detected. Therefore, it can be assumed that all three descriptor types are suitable for modeling the ABCB1 substrate/non-substrate classification problem. Independent of the descriptor type the global data set achieved an average

performance compared to all other data sets. However, considering the fact that 26 of all models exhibited a classification accuracy of higher than 80%, it can be reasonably suggested that all of these models might be very useful for external prediction of new compounds.

Assesing the "effectiveness" (external validaty) of RF-ABCB1 classification models using ROC curves.

In order to probe the effectiveness (generalizability) of the established models, we challenged the 27 classification models with an external test set. This external data set was provided by Boehringer Ingelheim and was previously used by us for RuleFit modeling (9). The external data set EXT consisted originally of 1436 compounds. However, for 13 compounds the ACORR descriptors could not be calculated due to software limitations and therefore this data set was reduced to 1423 molecules (120 substrates/1303 non-substrates) for all descriptor types to provide comparability. The classification performance for this EXT set is summarized in the following table by means of MCC and overall classification accuracy.

The confusion matrices for all the 27 models and their further numeric performance assessments for the EXT set are provided in supplementary information S3 at the end of this chapter. Supplementary information S4 provides an additional visualization of the prediction performance of EXT using ROC plots.

Table 3: Summary of the classification performance for the EXT set of the 27 models using MCC and acc as performance measures. The best performing models (assessed by MCC) are depicted in bold letters, whereas the poor performing models are depicted in italics.

	ACORR		MACCS		VSA	
model	**MCC**	**acc**	**MCC**	**acc**	**MCC**	**acc**
bi_div	0.56	0.94	0.32	0.92	0.36	0.92
bi_hom	0.14	0.38	0.32	0.68	0.34	0.68
bidivbihom	**0.79**	0.97	**0.77**	0.96	**0.80**	0.97
bimd	0.68	0.94	0.13	0.40	0.27	0.57
bimdlitnci	0.67	0.94	0.19	0.51	0.26	0.55
global	**0.77**	0.97	**0.74**	0.96	**0.71**	0.95
lit	*0.06*	*0.21*	*0.14*	*0.29*	*0.15*	0.31
litnci	0.22	0.53	0.14	0.38	0.18	0.45
nci	*0.08*	0.83	*0.03*	0.89	*-0.02*	0.89

From Table 3 it can be seen that the best model was retrieved with the merged BIA set (BIDIVBIHOM) on basis of the 32 VSA descriptors. This model returned a MCC of 0.80 and an overall classification accuracy of 0.97. Additionally, this model showed a sensitivity of 0.80 and a specificity of 0.99 (see also supplementary information S3.1). Furthermore, this data set gave also the best performance for the other two descriptor sets (ACORR:

MCC=0.79, acc=0.97; MACCS: MCC=0.77, acc=0.96). The worst performance in predicting the EXT data set was observed for the NCI set when using VSA descriptors. Here, a MCC=-0.02 was achieved. In general, the NCI and the LIT set showed unsatisfying performance when challenged with this external test set. Additionally, the BI_HOM set gave a rather poor performance when predicting the EXT set compared to its training performance. In conclusion, from Table 3 it can be summarized that the models built using BIA data sets, BI_DIV, to some extent also BI_HOM and their merged variant) as well as the global model outperform models that are constructed using the LIT, NCI or a combination of this two sets. The descriptors that contribute most to the model are presented in Supplementary Information S5. Furthermore, the ability to reliably predict the EXT set turns out to be dependent on the data set rather than the chemical descriptors used to describe the respective molecules.

These results are to some extent surprising than all models showed a rather satisfying performance in internal validation. The external prediction performance of the LIT and NCI set and to some extent also the BI_HOM set is indicative of an "efficacy-effectiveness"-gap. The failure to reliably predict the external test set can be most likely attributed to the fact that the LIT and NCI sets are substantially different with respect to their underlying chemistry compared to the BIA training sets and the BIA EXT set.

In order to account for these putative differences in chemistry, two different D2M measures were applied in order to estimate the applicability domain of the models.

The influence of post-processing methods that assess the applicability domain of RF-ABCB1 classification model performance

In an effort to estimate the applicability domain of the 27 models, the EXT set was subjected to the two D2M measures *range.tsh* and *dist.tsh*. Subsequently, those compounds suggested to be out-of-domain by the respective D2M measure, *range.tsh* or *dist.tsh* were removed from the EXT data set and the prediction performance was re-evaluated only on basis of those molecules classified as being "in-domain". This re-evaluation was carried out for all 27 models and for both D2M measures. The final results retrieved from this post-processing method are graphically summarized in Figure 2 and Figure 3. The bar charts in these two figures report the *change in MCC* upon applying either the *range.tsh* or the *dist.tsh* method. In both figures *change in MCC* is defined as the difference of the obtained MCC after applying the respective D2M measure (i.e. retrieved from model re-evaluation after removing those "out-of-domain" molecules from the EXT set) and the MCC using all molecules (i.e. MCC obtained from the prediction of EXT without applying a D2M measure) for the respective models.

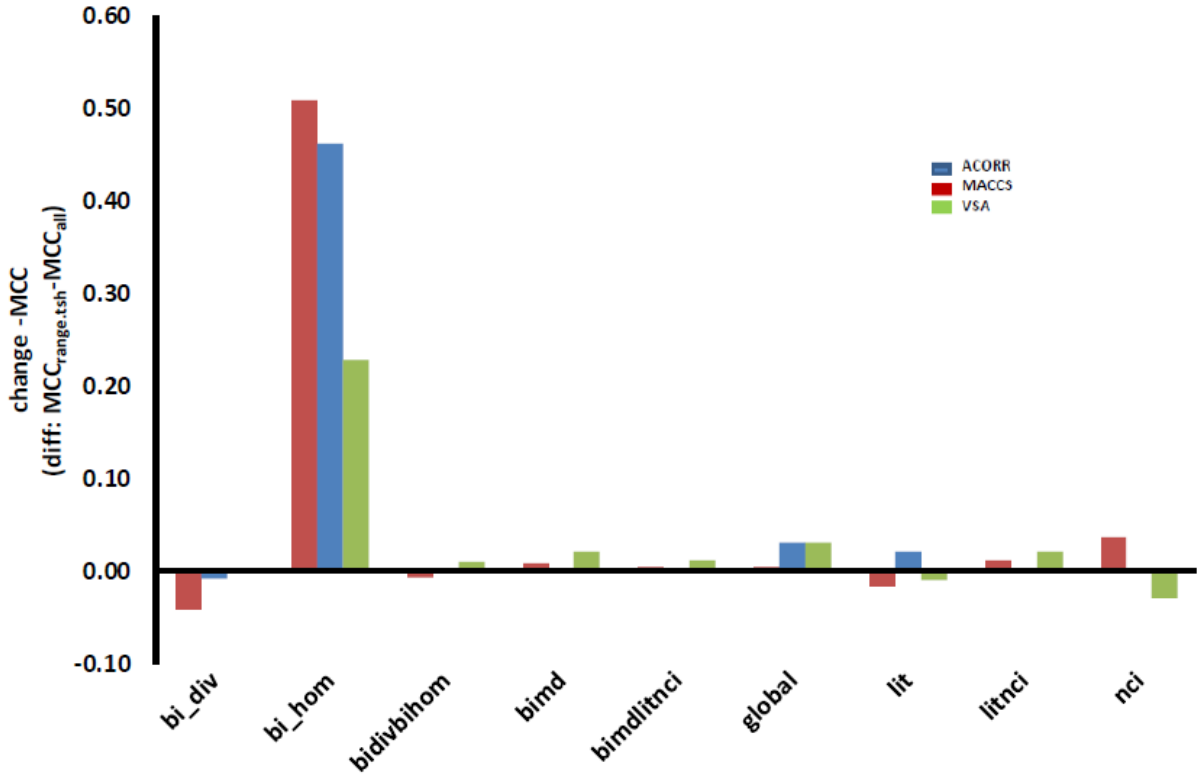


Figure 2: Influence of the application of range.tsh, a range-based D2M measure, on classification performance for the EXT set. Classification performance is defined as MCC. The x-axis displays the different training sets and the colours denote the three descriptor types. The y-axis displays the change in MCC that occurs upon applying range.tsh. It is calculated as $MCC_{range.tsh}$ - MCC_{all}

The results summarized in Figure 2 and Figure 3 must therefore be interpreted as follows: those models that exhibit positive values with respect to *change in MCC*, show an improvement in performance upon applying the respective D2M measure, whereas those models with negative *change in MCC* values show a decrease in performance.

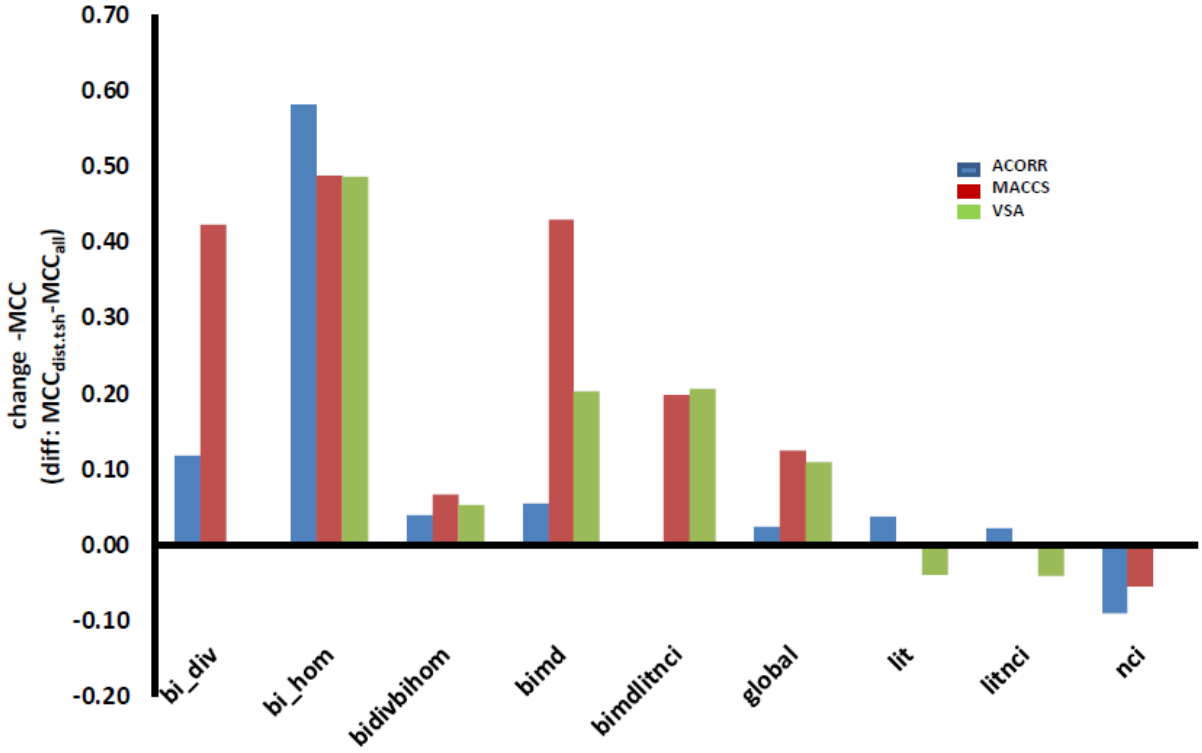


Figure 3: Influence of the application of dist.tsh, a distance-based AD measure, on classification performance for the EXT set. Classification performance is defined as MCC. The x-axis displays the different training sets and the colours denote the three descriptor types. The y-axis displays the change in MCC that occurs upon applying rang.tsh. It is calculated as $MCC_{dist.tsh}$ - MCC_{all}

When comparing Figure 2 and Figure 3 it can be seen that in general, the distance-based D2M measure, *dist.tsh* (shown in Figure 3), returns an improvement for all models based on BI training sets, whereas *range.tsh* (shown in Figure 2) is only capable of improving the performance of the BI_HOM model. The BI_HOM the weakest performance in predicting the EXT set.

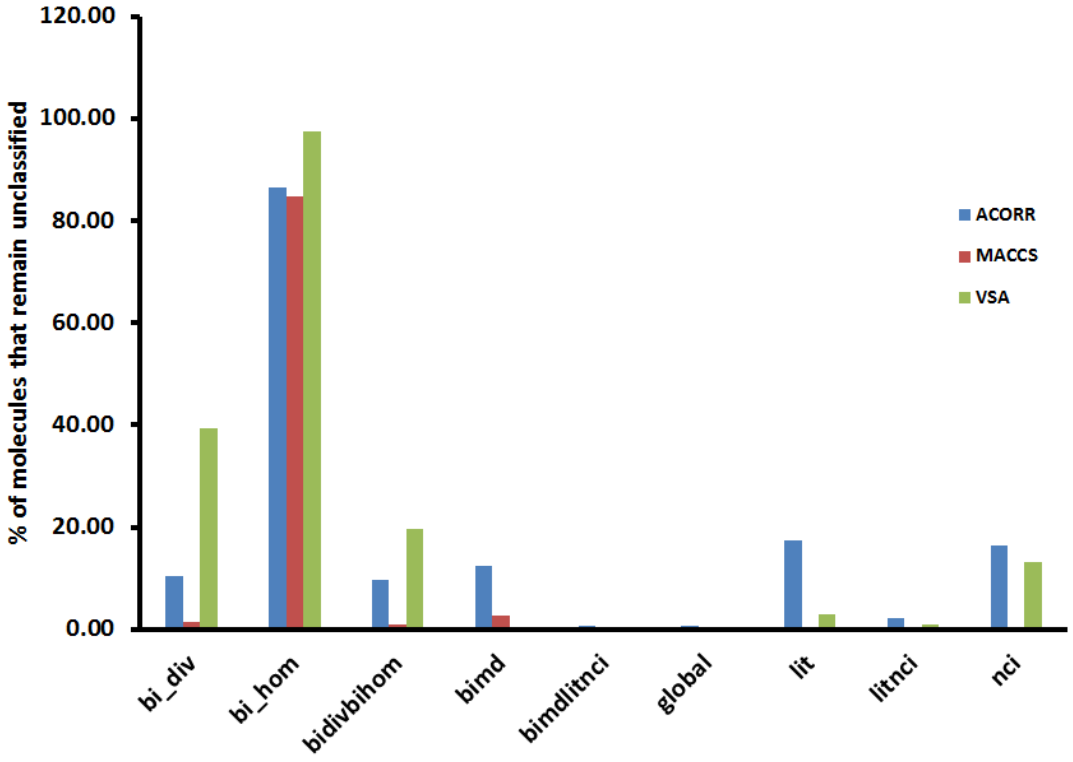


Figure 4: Percentage of molecules that remain unclassified or are considered to be out of domain per model after applying range.tsh. The x-axis denotes the different models and the colours encode the different descriptor sets.

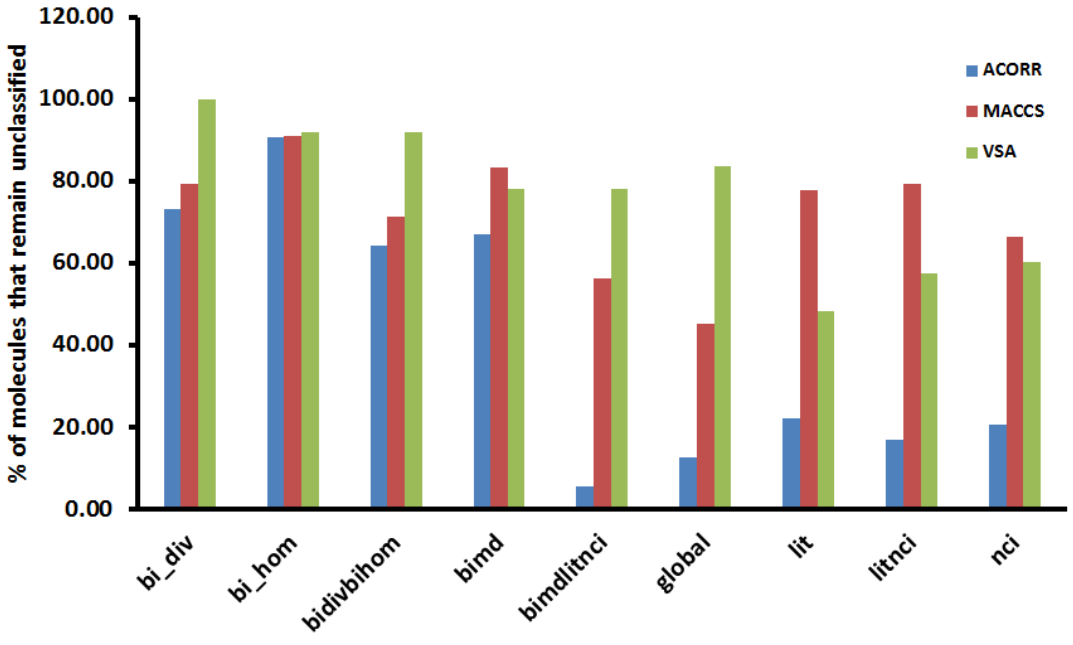


Figure 5: Percentage of molecules that remain unclassified or are considered to be out of domain per model after applying dist.tsh. The x-axis denotes the different models and the colours encode the different descriptor sets.

From Figure 4 and Figure 5 it can be seen that the application of *dist.tsh* leaves a substantial amount of unclassified molecules or in other words classifies a lot of molecules as being "out-of-domain". Contrary to that, the range-based D2M measure *range.tsh* classifies only a minority of molecules as being out-of-domain and therefore only few compounds are left

unclassified. The fact that *range.tsh* (Figure 4) considers only a minority of molecules "out-of-domain", while *dist.tsh* (Figure 5) categorizes many compounds to be "out-of-domain", suggests that *range.tsh* is too soft. Another interesting finding is that both D2M measures improve the performance of BI_HOM. BI_HOM is a very homogeneous data set that contains only 227 molecules. This suggests that models built on small data sets that only enclose a limited chemical space tend to profit most from a post-processing D2M measure, while the performance of models built on very diverse data sets remains almost unaffected. A similar observation was found recently by Weaver et al. [31]. Based on the fact, that the distance-based D2M, *dist.tsh*, outperforms, the simpler range-based method, *range.tsh*, with respect to *change in MCC*, the focus of the further discussion concentrates only on the performance and behaviour of *dist.tsh*.

Probably, most interesting is the finding that none of the two measures influences the performance of the LIT, NCI or the LITNCI set. This is of particular interest, because *dist.tsh* assigns 20% or up to 80% (depending on the descriptor set applied) of the EXT set as being "out-of-domain". A possible explanation for this is that the definition of ABCB1 substrates for the LIT and the NCI data set are based on different experiments. The LIT compounds comprise a collection of different assays (e.g. efflux assay or ATPase assay) used to determine ABCB1 substrates, whereas the NCI set is based on Pearson`s correlation coefficients for the cytotoxicity of a molecule and the ABCB1-mRNA content in 60 different cell lines for substrate determination. Contrary to that, the BIA sets (BIA_HOM, BIA_DIV, BIA_EXT) define substrates by comparing the cytotoxicity of a molecule in a transporter expressing cell lines and a transporter naïve cell line. If one now recalls the definition of the OECD for the AD of a QSAR model, which says:"*The applicability domain of a (Q)SAR model is the **response** and chemical structure **space** in which the model makes predictions with a given reliability*" [32], it becomes clear that the EXT set is based on a different response space than the training sets: LIT, NCI or LITNCI. This suggests that different pharmacological assays used to identify ABC-transporter substrates, might capture different aspects of the pharmacological effects of the compounds. Furthermore, it can be suggested that these assays are not directly comparable. This phenomenon was also recently analysed and critically discussed by Zdrazil et al. [33].

Conclusion

In summary, it can be concluded that the application of D2M measures can be useful for improving the performance of models that are based on small, homogeneous training sets. For models based on large, diverse data sets these post-processing techniques seem to be of less benefit. It has been shown that a distance-based measure is superior to a simple range-based measure. Last but not least, D2M measures can only capture differences in the chemistry of the training and the test set. Therefore, it is of utmost importance to make sure that the test set also captures the same pharmacological aspects (response) as the training set. In conclusion, the occurrence of "efficacy-effectiveness" gaps can have two reasons: training and test set are not compatible with respect to their chemistry (chemical space) or training and test set are not compatible with respect to the assessment of their pharmacology (response space). Both cases occurred in this study.

References

[1] G. Szakács, J. K. Paterson, J. a Ludwig, C. Booth-Genthe, and M. M. Gottesman, "Targeting multidrug resistance in cancer.," *Nature Reviews Drug Discovery*, vol. 5, no. 3, pp. 219–34, Mar. 2006.

[2] J. P. Gillet and M. M. Gottesman, "Mechanisms of multidrug resistance in cancer," *Methods Mol. Biol.*, vol. 596, pp. 47–76.

[3] A. Bodó, E. Bakos, F. Szeri, A. Váradi, and B. Sarkadi, "The role of multidrug transporters in drug availability, metabolism and toxicity," *Toxicology Letters*, vol. 140–141, pp. 133–143.

[4] B. Sarkadi, L. Homolya, G. Szakács, and A. Váradi, "Human multidrug resistance ABCB and ABCG transporters: participation in a chemoimmunity defense system," *Physiological Reviews*, vol. 86, no. 4, pp. 1179–1236.

[5] S. G. Aller, J. Yu, A. Ward, Y. Weng, S. Chittaboina, R. Zhuo, P. M. Harrell, Y. T. Trinh, Q. Zhang, I. L. Urbatsch, and G. Chang, "Structure of P-glycoprotein reveals a molecular basis for poly-specific drug binding," *Science*, vol. 323, pp. 1718–1722.

[6] J. W. Polli, S. a Wring, J. E. Humphreys, L. Huang, J. B. Morgan, L. O. Webster, and C. S. Serabjit-Singh, "Rational use of in vitro P-glycoprotein assays in drug discovery.," *The Journal of Pharmacology and Experimental Therapeutics*, vol. 299, no. 2, pp. 620–8, Nov. 2001.

[7] M. A. Demel, R. Schwaha, O. Krämer, P. Ettmayer, E. E. Haaksma, and G. F. Ecker, "In silico prediction of substrate properties for ABC-multidrug transporters.," *Expert Opinion on Drug Metabolism & Toxicology*, vol. 4, no. 9, pp. 1167–80, Sep. 2008.

[8] M. A. Demel, O. Krämer, P. Ettmayer, E. E. J. Haaksma, and G. F. Ecker, "Predicting ligand interactions with ABC transporters in ADME," *Chemistry & Biodiversity*, vol. 6, no. 11, pp. 1960–1969.

[9] M. A. Demel, O. Kraemer, P. Ettmayer, E. Haaksma, and G. F. Ecker, "Ensemble Rule-Based Classification of Substrates of the Human ABC-Transporter ABCB1 Using Simple Physicochemical Descriptors," *Molecular Informatics*, vol. 29, no. 3, pp. 233–242, Mar. 2010.

[10] A. Golbraikh and A. Tropsha, "Beware of q2!," *Journal of Molecular Graphics & Modelling*, vol. 20, no. 4, pp. 269–276.

[11] H.-G. Eichler, E. Abadie, A. Breckenridge, B. Flamion, L. L. Gustafsson, H. Leufkens, M. Rowland, C. K. Schneider, and B. Bloechl-Daum, "Bridging the efficacy-effectiveness gap: a regulator's perspective on addressing variability of drug response.," *Nature reviews. Drug discovery*, vol. 10, no. 7, pp. 495–506, Jul. 2011.

[12] G. Szakacs, J. P. Annereau, S. Lababidi, U. Shankavaram, A. Arciello, K. J. Bussey, W. Reinhold, Y. Guo, G. D. Kruh, M. Reimers, J. N. Weinstein, and M. M. Gottesman, "Predicting drug sensitivity and resistance: profiling ABC transporter genes in cancer cells," *Cancer Cell*, vol. 6, pp. 129–137.

[13] M. Demel, A. G. . Janecek, W. Gansterer, and G. F. Ecker, “Comparison of Contemporary Feature Selection Algorithms: Application to the Classification of ABC-Transporter Substrates,” no. 10, pp. 1087–1091.

[14] A. Seelig, “A general pattern for substrate recognition by P-glycoprotein,” *Eur. J. Biochem.*, vol. 251, pp. 252–261.

[15] A. Seelig and E. Landwojtowicz, “Structure-activity relationship of P-glycoprotein substrates and modifiers,” *Eur J Pharm Sci*, vol. 12, pp. 31–40.

[16] V. K. Gombar, J. W. Polli, J. E. Humphreys, S. A. Wring, and C. S. Serabjit-Singh, “Predicting P-glycoprotein substrates by a quantitative structure-activity relationship model,” *J Pharm Sci*, vol. 93, pp. 957–968.

[17] J. Huang, G. Ma, I. Muhammad, and Y. Cheng, “Identifying P-glycoprotein substrates using a support vector machine optimized by a particle swarm,” *J Chem Inf Model*, vol. 47, pp. 1638–1647.

[18] R. Didziapetris, P. Japertas, A. Avdeef, and A. Petrauskas, “Classification analysis of P-glycoprotein substrate specificity,” *J Drug Target*, vol. 11, pp. 391–406.

[19] A. Garrigues, N. Loiseau, M. Delaforge, J. Ferte, M. Garrigos, F. Andre, and S. Orlowski, “Characterization of two pharmacophores on the multidrug transporter P-glycoprotein,” *Mol. Pharmacol.*, vol. 62, pp. 1288–1298.

[20] P. Labute, “A widely applicable set of descriptors,” *Journal of Molecular Graphics & Modelling*, vol. 18, no. 4–5, pp. 464–477.

[21] L. Breiman, “Random Forests,” pp. 1–35, 1999.

[22] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, “Random forest: a classification and regression tool for compound classification and QSAR modeling,” *Journal of Chemical Information and Computer Sciences*, vol. 43, no. 6, pp. 1947–1958.

[23] F. Cheng, J. Shen, Y. Yu, W. Li, G. Liu, P. W. Lee, and Y. Tang, “In silico prediction of Tetrahymena pyriformis toxicity for diverse industrial chemicals with substructure pattern recognition and machine learning methods,” *Chemosphere*, vol. 82, no. 11, pp. 1636–1643.

[24] R. Schwaha and G. F. Ecker, “Use of shape similarities for the classification of P-glycoprotein substrates and nonsubstrates,” *Future Medicinal Chemistry*, vol. 3, no. 9, pp. 1117–1128, 2011.

[25] B. Chen, R. P. Sheridan, V. Hornak, and J. H. Voigt, “Comparison of random forest and Pipeline Pilot Naive Bayes in prospective QSAR predictions,” *Journal of Chemical Information and Modeling*.

[26] J. A. Hanley, “Receiver operating characteristic (ROC) methodology: the state of the art,” *Critical Reviews in Diagnostic Imaging*, vol. 29, no. 3, pp. 307–335.

[27] J. Jaworska, N. Nikolova-Jeliazkova, and T. Aldenberg, "QSAR applicabilty domain estimation by projection of the training set descriptor space: a review," *Alternatives to Laboratory Animals: ATLA*, vol. 33, no. 5, pp. 445–459.

[28] I. V. Tetko, I. Sushko, A. K. Pandey, H. Zhu, A. Tropsha, E. Papa, T. Oberg, R. Todeschini, D. Fourches, and A. Varnek, "Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection," *Journal of Chemical Information and Modeling*, vol. 48, no. 9, pp. 1733–1746.

[29] I. Sushko, S. Novotarskyi, R. Körner, A. K. Pandey, A. Cherkasov, J. Li, P. Gramatica, K. Hansen, T. Schroeter, K.-R. Müller, L. Xi, H. Liu, X. Yao, T. Öberg, F. Hormozdiari, P. Dao, C. Sahinalp, R. Todeschini, P. Polishchuk, A. Artemenko, V. Kuz'min, T. M. Martin, D. M. Young, D. Fourches, E. Muratov, A. Tropsha, I. Baskin, D. Horvath, G. Marcou, C. Muller, A. Varnek, V. V. Prokopenko, and I. V. Tetko, "Applicability domains for classification problems: Benchmarking of distance to models for Ames mutagenicity set," *Journal of Chemical Information and Modeling*, vol. 50, no. 12, pp. 2094–2111.

[30] A. Tropsha and A. Golbraikh, "Predictive QSAR modeling workflow, model applicability domains, and virtual screening," *Current Pharmaceutical Design*, vol. 13, no. 34, pp. 3494–3504.

[31] S. Weaver and M. P. Gleeson, "The importance of the domain of applicability in QSAR modeling.," *Journal of Molecular Graphics & Modelling*, vol. 26, no. 8, pp. 1315–26, Jun. 2008.

[32] P. A. B. OECD WORKING PARTY ON CHEMICALS, "OECD Environment Health and Safety Publications Series on Testing and Assessment No . 69 GUIDANCE DOCUMENT ON THE VALIDATION OF (QUANTITATIVE) STRUCTURE-ACTIVITY RELATIONSHIP [(Q) SAR] MODELS Environment Directorate," 2007.

[33] B. Zdrazil, M. Pinto, P. Vasanthanathan, A. J. Williams, L. Z. Balderud, O. Engkvist, C. Chichester, A. Hersey, J. P. Overington, and G. F. Ecker, "Annotating Human P-Glycoprotein Bioassay Data," *Molecular Informatics*, p. n/a–n/a, Aug. 2012.

Author Contribution

The BIA sets were provided by Boehringer Ingelheim Austria (Dr. Oliver Krämer). Dr. Krämer also provided the external test set. The BIA team (Dr. Haaksma, Dr. Ettmayer, Dr. Krämer) assisted in the design of computational experiments. The choice of using range.tsh and dist.tsh was done by me and Dr. Krämer. I wrote all the R-scripts and performed all the calculations. The results were evaluated by me and Dr. Krämer. Prof. Ecker supervised this project.

Supplementary information

Supplementary information S1:

R-source code for dist.tsh:

The function expects three input parameters (the training set, the test set, the parameter Z) and returns a data matrix, with three columns (the compound ID of the test compounds, the determined distance, the dichotomized label for AD). NOTE: the function requires that the R *proxy()* package is properly installed and loaded. *proxy()* can be obtained from CRAN.

```
require(proxy)
dist.tsh=function(x,DD,Z=0.5){
  knn.DD.train=0
  tmp=0
  knn.train=0
  knn.test=0
  DD.train=as.matrix(dist(x[,2:(length(x)-1)]))


  for (i in 1:dim(DD.train)[2]){
    tmp=sort(DD.train[i,])
    knn.train[i]=tmp[2]
  }

   mean.DD.train=mean(knn.train)
   sd.DD.train=sd(knn.train)

      ####### determine distance threshold
  DD.dist.tsh=mean.DD.train+Z*sd.DD.train

#################
#################
#################

mat=as.matrix(dist(x[,1:length(x)],DD[,1:length(DD)], method="euclidean",
scale=T))

    for (i in 1:dim(DD)[1]){
               tmp=sort(mat[,i])
                  knn.test[i]=tmp[2]
    }
  #################
# dichotomize distance
    DD.passed=knn.test<DD.dist.tsh

  for(i in 1:length(DD.passed)){
    if(DD.passed[i]=="FALSE")
      DD.passed[i]="outofDOM"
    if(DD.passed[i]=="TRUE")
      DD.passed[i]="inDOM"
  }
# generate return argument and return

  DD.dom=as.data.frame(cbind(dist=knn.test,dist.tsh=DD.passed))
return(DD.dom)
}
```

Supplementary information S2:

R-source code for range.tsh:

The function expects three input parameters (the training set, the test set, the parameter cut) and returns a data matrix, with three columns. NOTE: input parameter cut defaults to 0.1; i.e. in case that

a test compound violates the margin to the training compounds, in more than 10% of the cases, it will be considered to be “outofDOM”.

```
range.dist=function(train,test, cut=0.1){
  min.train=vector()
  max.train=vector()
  test.min=vector()
  test.max=vector()
  tmp=vector()
  train=train[,-1]
  train=train[,-dim(train)[2]]
  test=test[,-dim(test)[2]]
  dist.based=vector()
  range=matrix(,dim(test)[1],dim(test)[2])

 for(j in 1:dim(test)[1]){
    for(i in 1:length(test)){
 # determine min(), max() values for each descriptor in
 # the training matrix
    min.train[i]=min(train[,i])
             max.train[i]=max(train[,i])
 # compare min,max-values to the test set
      if(test[j,i]>=min.train[i] && test[j,i]<=max.train[i]) {
          range[j,i]="TRUE"}
        else{
           range[j,i]="FALSE"}
    }
  }
 # count occurences
  for(k in 1:dim(range)[1]){

tmp[k]=sum(range[k,1:dim(range)[2]]=="FALSE")/sum(range[k,1:dim(range)[2]]=
="TRUE")
  }
 # label test IDs according to cut=0.1
  for(k in 1:dim(range)[1]){
   if(tmp[k]<=cut)
    dist.based[k]="inDOM"
   else
    dist.based[k]="outofDOM"
  }

  return(dist.based)
}
```

Supplementary information S3:

Confusion matrices and performance measures for the prediction of the external set EXT for all models.

Table S3.1: Prediction of EXT. Performance without considering an AD measure

D2M	desc type	model	tp	tn	fp	fn	MCC	acc	sens	spec	pract	prinact
none	ACORR	bi_div	54	1288	16	66	0.56	0.94	0.45	0.99	0.77	0.95
none	ACORR	bi_hom	107	438	869	12	0.14	0.38	0.90	0.34	0.11	0.97
none	ACORR	bidivbihom	95	1282	22	24	0.79	0.97	0.80	0.98	0.81	0.98

none	ACORR	bimd	97	1246	58	22	0.68	0.94	0.82	0.96	0.63	0.98
none	ACORR	bimdlitnci	99	1236	68	20	0.67	0.94	0.83	0.95	0.59	0.98
none	ACORR	global	94	1280	24	25	0.77	0.97	0.79	0.98	0.80	0.98
none	ACORR	lit	111	188	1117	8	0.06	0.21	0.93	0.14	0.09	0.96
none	ACORR	litnci	108	646	661	11	0.22	0.53	0.91	0.49	0.14	0.98
none	ACORR	nci	24	1161	143	97	0.08	0.83	0.20	0.89	0.14	0.92

D2M	desc type	model	tp	tn	fp	fn	MCC	acc	sens	spec	pract	prinact
none	MACCS	bi_div	32	1271	33	87	0.32	0.92	0.27	0.97	0.49	0.94
none	MACCS	bi_hom	107	862	445	12	0.32	0.68	0.90	0.66	0.19	0.99
none	MACCS	bidivbihom	98	1273	31	21	0.77	0.96	0.82	0.98	0.76	0.98
none	MACCS	bimd	104	460	844	15	0.13	0.40	0.87	0.35	0.11	0.97
none	MACCS	bimdlitnci	104	619	685	15	0.19	0.51	0.87	0.47	0.13	0.98
none	MACCS	global	99	1263	42	20	0.74	0.96	0.83	0.97	0.70	0.98
none	MACCS	lit	117	299	1006	2	0.14	0.29	0.98	0.23	0.10	0.99
none	MACCS	litnci	108	432	874	11	0.14	0.38	0.91	0.33	0.11	0.98
none	MACCS	nci	6	1266	38	113	0.03	0.89	0.05	0.97	0.14	0.92

D2M	desc type	model	tp	tn	fp	fn	MCC	acc	sens	spec	pract	prinact
none	VSA	bi_div	1	1303	1	118	0.06	0.92	0.01	1.00	0.50	0.92
none	VSA	bi_hom	111	860	446	8	0.34	0.68	0.93	0.66	0.20	0.99
none	VSA	bidivbihom	95	1286	18	24	0.80	0.97	0.80	0.99	0.84	0.98
none	VSA	bimd	113	697	608	6	0.27	0.57	0.95	0.53	0.16	0.99
none	VSA	bimdlitnci	114	667	640	5	0.26	0.55	0.96	0.51	0.15	0.99
none	VSA	global	102	1246	58	17	0.71	0.95	0.86	0.96	0.64	0.99
none	VSA	lit	116	326	978	3	0.15	0.31	0.97	0.25	0.11	0.99
none	VSA	litnci	108	529	775	11	0.18	0.45	0.91	0.41	0.12	0.98
none	VSA	nci	2	1271	33	117	-0.02	0.89	0.02	0.97	0.06	0.92

Table S3.2: Prediction of EXT. Performance after applying range.tsh; only considering molecules to be inside the applicability domain

D2M	desc type	model	tp	tn	fp	fn	MCC	acc	sens	spec	pract	princact	% act. uncl.	% inact. uncl.	% uncl. total.
range.tsh	ACORR	bi_div	42	1157	15	61	0.52	0.94	0.41	0.99	0.74	0.95	14.17	10.12	10.46
range.tsh	ACORR	bi_hom	82	76	25	10	0.65	0.82	0.89	0.75	0.77	0.88	23.33	92.25	86.45
range.tsh	ACORR	bidivbihom	87	1156	21	23	0.78	0.97	0.79	0.98	0.81	0.98	8.33	9.74	9.62
range.tsh	ACORR	bimd	91	1084	53	19	0.69	0.94	0.83	0.95	0.63	0.98	8.33	12.81	12.43
range.tsh	ACORR	bimdlitnci	99	1227	68	19	0.67	0.94	0.84	0.95	0.59	0.98	1.67	0.69	0.77
range.tsh	ACORR	global	94	1271	24	24	0.78	0.97	0.80	0.98	0.80	0.98	1.67	0.69	0.77
range.tsh	ACORR	lit	97	88	986	4	0.04	0.16	0.96	0.08	0.09	0.96	15.83	17.64	17.49
range.tsh	ACORR	litnci	108	643	633	10	0.23	0.54	0.92	0.50	0.15	0.98	1.67	2.15	2.11
range.tsh	ACORR	nci	21	995	86	89	0.11	0.85	0.19	0.92	0.20	0.92	8.33	17.10	16.36

D2M	desc type	model	tp	tn	fp	fn	MCC	acc	sens	spec	pract	princact	% act. uncl.	% inact. uncl.	% uncl. total.
range.tsh	MACCS	bi_div	30	1256	33	84	0.31	0.92	0.26	0.97	0.48	0.94	5.00	1.15	1.47
range.tsh	MACCS	bi_hom	91	101	14	10	0.78	0.89	0.90	0.88	0.87	0.91	15.83	91.18	84.83
range.tsh	MACCS	bidivbihom	96	1264	31	21	0.77	0.96	0.82	0.98	0.76	0.98	2.50	0.69	0.84
range.tsh	MACCS	bimd	102	447	823	15	0.13	0.40	0.87	0.35	0.11	0.97	2.50	2.61	2.60
range.tsh	MACCS	bimdlitnci	104	619	685	15	0.19	0.51	0.87	0.47	0.13	0.98	0.83	0.00	0.07
range.tsh	MACCS	global	99	1263	42	20	0.74	0.96	0.83	0.97	0.70	0.98	0.83	-0.08	0.00
range.tsh	MACCS	lit	114	299	1006	2	0.14	0.29	0.98	0.23	0.10	0.99	3.33	-0.08	0.21
range.tsh	MACCS	litnci	108	432	874	11	0.14	0.38	0.91	0.33	0.11	0.98	0.83	-0.15	-0.07
range.tsh	MACCS	nci	6	1266	38	113	0.03	0.89	0.05	0.97	0.14	0.92	0.83	0.00	0.07

D2M	desc type	model	tp	tn	fp	fn	MCC	acc	sens	spec	pract	princact	% act. uncl.	% inact. uncl.	% uncl. total.
range.tsh	VSA	bi_div	0	819	0	43	NA	0.95	0.00	1.00	NA	0.95	64.17	37.19	39.47
range.tsh	VSA	bi_hom	27	3	5	0	0.56	0.86	1.00	0.38	0.84	1.00	77.50	99.39	97.54
range.tsh	VSA	bidivbihom	85	1024	17	18	0.81	0.97	0.83	0.98	0.83	0.98	14.17	20.17	19.66
range.tsh	VSA	bimd	113	697	608	6	0.27	0.57	0.95	0.53	0.16	0.99	0.83	0.08	0.00

range.tsh	VSA	bimdlitnci	114	667	640	5	0.26	0.55	0.96	0.51	0.15	0.99	0.83	0.23	0.14
range.tsh	VSA	global	102	1246	58	17	0.71	0.95	0.86	0.96	0.64	0.99	0.83	0.00	0.07
range.tsh	VSA	lit	114	312	953	3	0.15	0.31	0.97	0.25	0.11	0.99	2.50	2.99	2.95
range.tsh	VSA	litnci	108	522	770	11	0.18	0.45	0.91	0.40	0.12	0.98	0.83	0.92	0.91
range.tsh	VSA	nci	2	1097	27	112	0.01	0.89	0.02	0.98	0.07	0.91	5.00	13.80	13.06

Table S3.3: Performance after applying dist.tsh; only considering molecules to be inside the applicability domain

D2M	desc type	model	tp	tn	fp	fn	MCC	acc	sens	spec	pract	prinact	% act. uncl.	% inact. uncl.	% uncl. total.
dist.tsh	ACORR	bi_div	8	366	1	7	0.68	0.98	0.53	1.00	0.89	0.98	87.50	71.86	73.17
dist.tsh	ACORR	bi_hom	46	68	11	7	0.72	0.86	0.87	0.86	0.81	0.91	55.83	93.94	90.73
dist.tsh	ACORR	bidivbihom	60	427	9	12	0.83	0.96	0.83	0.98	0.87	0.97	40.00	66.56	64.33
dist.tsh	ACORR	bimd	83	343	29	14	0.74	0.91	0.86	0.92	0.74	0.96	19.17	71.47	67.06
dist.tsh	ACORR	bimdlitnci	99	1160	68	19	0.67	0.94	0.84	0.94	0.59	0.98	1.67	5.83	5.48
dist.tsh	ACORR	global	91	1114	22	19	0.80	0.97	0.83	0.98	0.81	0.98	8.33	12.88	12.50
dist.tsh	ACORR	lit	105	111	892	1	0.10	0.19	0.99	0.11	0.11	0.99	11.67	23.08	22.12
dist.tsh	ACORR	litnci	104	483	595	3	0.24	0.50	0.97	0.45	0.15	0.99	10.83	17.33	16.78
dist.tsh	ACORR	nci	5	956	123	48	0.01	0.85	0.09	0.89	0.04	0.95	55.83	17.25	20.51
D2M	**desc type**	**model**	**tp**	**tn**	**fp**	**fn**	**MCC**	**acc**	**sens**	**spec**	**pract**	**prinact**	**% act. uncl.**	**% inact. uncl.**	**% uncl. total.**
dist.tsh	MACCS	bi_div	3	289	1	1	0.75	0.99	0.75	1.00	0.75	1.00	96.67	77.76	79.35
dist.tsh	MACCS	bi_hom	37	79	6	5	0.81	0.91	0.88	0.93	0.86	0.94	65.00	93.48	91.08
dist.tsh	MACCS	bidivbihom	35	361	6	6	0.84	0.97	0.85	0.98	0.85	0.98	65.83	71.86	71.35
dist.tsh	MACCS	bimd	49	140	42	7	0.56	0.79	0.88	0.77	0.54	0.95	53.33	86.04	83.29
dist.tsh	MACCS	bimdlitnci	92	337	181	14	0.39	0.69	0.87	0.65	0.34	0.96	11.67	60.28	56.18
dist.tsh	MACCS	global	77	682	10	11	0.86	0.97	0.88	0.99	0.89	0.98	26.67	46.93	45.22
dist.tsh	MACCS	lit	0	194	122	0	NA	0.61	NA	0.61	0.00	1.00	100.00	75.77	77.81
dist.tsh	MACCS	litnci	0	186	109	0	NA	0.63	NA	0.63	0.00	1.00	100.00	77.38	79.28

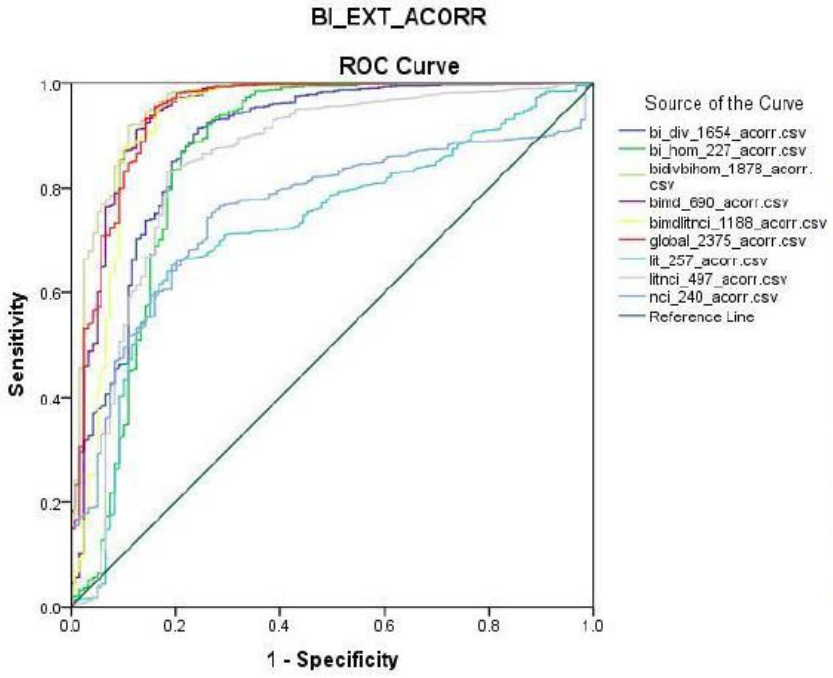
dist.tsh	MACCS	nci	0	456	19	5	0.02	0.95	0.00	0.96	0.00	0.99	95.83	63.57	66.29

D2M	desc type	model	tp	tn	fp	fn	MCC	acc	sens	spec	pract	prinact	% act. uncl.	% inact. uncl.	% uncl. total.
dist.tsh	VSA	bi_div	0	1	0	0	NA	1.00	NA	1.00	NA	1.00	100.00	99.92	99.93
dist.tsh	VSA	bi_hom	42	62	7	3	0.82	0.91	0.93	0.90	0.86	0.95	62.50	94.71	91.99
dist.tsh	VSA	bidivbihom	42	63	6	2	0.86	0.93	0.95	0.91	0.88	0.97	63.33	94.71	92.06
dist.tsh	VSA	bimd	94	117	98	5	0.47	0.67	0.95	0.54	0.49	0.96	17.50	83.51	77.95
dist.tsh	VSA	bimdlitnci	94	115	99	5	0.47	0.67	0.95	0.54	0.49	0.96	17.50	83.59	78.02
dist.tsh	VSA	global	79	133	15	5	0.82	0.91	0.94	0.90	0.84	0.96	30.00	88.65	83.71
dist.tsh	VSA	lit	21	210	505	0	0.11	0.31	1.00	0.29	0.04	1.00	82.50	45.17	48.31
dist.tsh	VSA	litnci	9	337	261	0	0.14	0.57	1.00	0.56	0.03	1.00	92.50	54.14	57.37
dist.tsh	VSA	nci	0	552	12	4	0.01	0.97	0.00	0.98	0.00	0.99	96.67	56.75	60.11

Supplementary information S4:

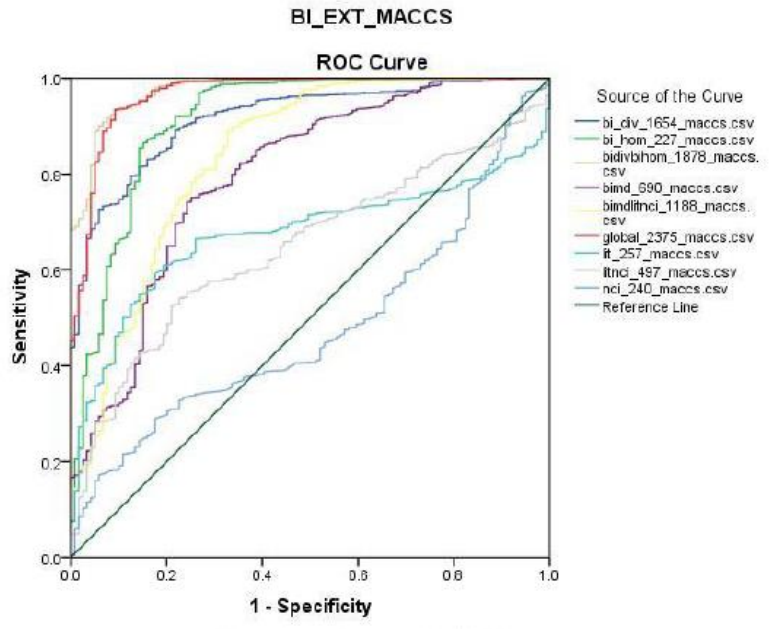
Visualization of the prediction performance for the EXT set by the different models based on the different descriptor sets using ROC plots and ROC-AUC as performance measure:

Figure S4.1: ROC plot for the prediction of EXT using ACORR descriptors



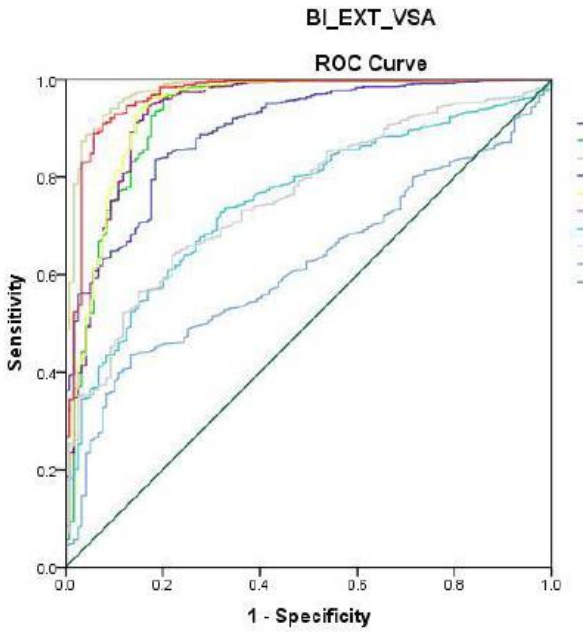
ACORR	
Model	ROC AUC
bi_div	0.89
bi_hom	0.86
bidivbihom	0.96
bimd	0.94
bimdlitnci	0.93
global	0.94
lit	0.72
litnci	0.84
nci	0.76

Figure S4.2: ROC plot for the prediction of EXT using MACCS descriptors



MACCS	
Model	ROC AUC
bi_div	0.92
bi_hom	0.91
bidivbihom	0.98
bimd	0.80
bimdlitnci	0.84
global	0.97
lit	0.67
litnci	0.65
nci	0.48

Figure S4.3: ROC plot for the prediction of EXT using VSA descriptors



VSA	
Model	**ROC AUC**
bi_div	**0.90**
bi_hom	**0.93**
bidivbihom	**0.98**
bimd	**0.93**
bimdlitnci	**0.94**
global	**0.97**
lit	**0.75**
litnci	**0.76**
nci	**0.62**

Supplementary Information S5: Variable Importances of the „global" model retrieved from the RF model based on MACCS-count descriptors

SUBSTRATES			NON-SUBSTRATES		
ID	**def.**	**RF. Imp.**	**ID**	**def.**	**RF. Imp.**
MACCS165	# RINGS	1.00	**MACCS62**	# N=O	1
MACCS113	# Onot%A%A	0.77	**MACCS91**	# OC(N)C	0.73
MACCS137	# heterocycle	0.43	**MACCS128**	# ACH2AAACH2A	0.97
MACCS111	# NACH2A	0.50	**MACCS105**	# A$A($A)$A	0.51
MACCS131	# QH > 1	0.50	**MACCS104**	# QHACH2A	0.50
MACCS97	# NAAAO	0.46	**MACCS129**	# ACH2AACH2A	0.46
MACCS124	# QQ	0.59	**MACCS157**	# C-O	0.30
MACCS151	# NH	0.51	**MACCS160**	# CH3	0.37
MACCS77	# NAN	0.26	**MACCS80**	NAAAN	0.34
MACCS155	# A!CH2!A	0.52	**MACCS117**	NAO	0.17
MACCS86	# CH2QCH2	0.36	**MACCS94**	QN	0.16
MACCS153	# QCH2A	0.38	**MACCS135**	# Nnot%A%A	0.16
MACCS72	# OAAO	0.36	**MACCS154**	# C=O	0.16
MACCS156	# NA(A)A	0.29	**MACCS143**	# A$A!O	0.12
MACCS98	# QAAAAA@1	0.25	**MACCS164**	# O	0.06

Chapter 14: Manuscript II - Application of Molecular Substructure-Similarity - guided Fingerprint SIBAR (MOSS-FP-SIBAR) Descriptors for the Classification of ABCB1-Substrates/Non-Substrates

Abstract

The similarity principle, which states that molecules that are similar with respect to their chemical structure are also likely to exhibit similar pharmacological effects, is well known in the field of medicinal chemistry and is widely applied in chemoinformatics. Recently, it was also applied in order to develop molecular descriptors as input for QSAR models. These new descriptors were termed similarity-based (SIBAR) descriptors and are derived from similarity calculations of the data set of interest towards a pre-defined reference set. These similarity values are consecutively used as inputs for predictive regression or classification modelling. Here, the SIBAR concept is further expanded. The molecular substructure-guided fingerprint SIBAR (MoSS-FP-SIBAR) descriptors outlined in this work differ from the recently introduced SIBAR descriptors with respect to the selection of the reference set, the chemical space and also with respect to the similarity metric that is applied. MoSS-FP-SIBAR is applied to the ABCB1 substrate/non-substrate classification problem. A random Forests ensemble classifier is used to compare MoSS-FP-SIBAR descriptors to classical fingerprint-based models and also to SIBAR descriptors based on a randomly selected reference set. It is demonstrated that, MoSS-FP-SIBAR descriptors show a similar performance to a fingerprint-based model and also harbour attractive features that render it highly suitable for interpretation purposes.

Introduction

Molecules that are similar with respect to their chemical structure tend to have similar biological activity. This principle is commonly applied in computational life sciences, especially in bioinformatics and also in chemoinformatics. In bioinformatics highly similar protein sequences are considered to reflect proteins with similar structures and similar functions. The application of the similarity principle in bioinformatics was the basis for phylogenetic comparisons and also homology modelling (Blundell, Jhoti, & Abell, 2002; Browne, North, & Phillips, 1969; Greer, 1980; Perutz, Kendrew, & Watson, 1965). In chemoinformatics, this common perception is additionally validated by a long list of historical examples (e.g. β-lactams are considered to exert antibacterial activity, NO-containing molecules are susceptible to elicit vasodilatating effects and phenyl-ethyl-amines are notorious for their CNS activities) and is the central driving force for the use of similarity-based methods (e.g. virtual screening, pharmacophore modelling) by computational chemists (Chen et al., 2009; Kubinyi, 1998; Martin, Kofron, & Traphagen, 2002). These similarity-based methods often make use of pairwise-similarity calculations which typically return a *n x n*-matrix for a set of *n* molecules. Furthermore, these methods usually make use of structural fingerprints, chemical descriptors or 3D-shapes to define the space in which the similarity calculation is performed. For binary fingerprints, the well-known Tanimoto coefficient is a commonly used similarity measure, whereas for continuous descriptor values, the Euclidean distance is usually applied as distance measure. Recently, the basic applicability for using

similarity values as descriptor values in the context of predictive modelling was demonstrated. The rationale for this approach is that similarity values, which are used as input for predictive modelling serve as substitute for the commonly used (calculated) molecular descriptor values. This method was termed SIBAR, which is short for similarity-based SAR. Briefly, the calculation of SIBAR descriptors in their original implementation is as follows (reviewed in Schwaha et al. (Schwaha & Ecker, 2008)): (i) selection of a reference set, (ii) calculation of chemical descriptors for both the references set as well as the data set of interest, (iii) calculation of Euclidean distances as distance measure between the calculated descriptors of the data set and the reference set, (iv) submission of these Euclidean distances to predictive modelling algorithms such as PLS, ANNs, CGP-NN or RF. SIBAR proved to be a useful tool for early ADME profiling, when applied to a data set of 131 propafenone-type ABCB1 inhibitors and 20 molecules showing intestinal absorption as pharmacological endpoint (Klein, Kaiser, Kopp, Chiba, & Ecker, 2002). Furthermore, it was also used in three other studies to model a larger set of more than 400 ABCB1 inhibitors, ABCB1 substrates and non-substrates and also a set of hERG inhibitors/non-inhibitors (Schwaha & Ecker, 2009; Thai & Ecker, 2009; Zdrazil, Kaiser, Kopp, Chiba, & Ecker, 2007). These studies extended the initial SIBAR concept by sampling the reference set from the data set of known actives and inactives and thereby tailoring it towards the biological problem. It is the primary objective of this study to further develop the concept of SIBAR descriptors previously described by Klein et al. (Klein et al., 2002). The extension presented here, differs with respect to the selection of references molecules from the known actives and inactives (i), the description of the similarity space (ii) and also the applied distance measure (iii). (i) In this study reference molecules are sampled from the data set of interest according to discriminatory substructures that are dominant in one class (*Mo*lecular *S*ub*S*tructure search - MoSS) – thereby also acknowledging the medicinal chemistry of the data set, (ii) molecules are described by various types of structural fingerprints (FP), and the (iii) Tanimoto coefficient in the range of [0,1] is used to compute similarities between the data set and the reference molecules. In summary, the approach is termed MoSS-FP-SIBAR, The feasibility of the MoSS-FP-SIBAR approach is critically appraised herein utilising a public available data set of ABCB1 substrates/non-substrates.

Material and Methods

Data Set

The data set of ABCB1 substrates and ABCB1 non-substrates was obtained from the publication of Poongavanam et al. (Poongavanam, Haider, & Ecker, 2012). It contains a compilation of 257 compounds collected from the literature and a set of 227 molecules extracted from the NCI60 drug sensitivity screen. In total this data set consists of 484 compounds and includes 243 substrates and 241 non-substrates. The data set was downloaded in .sdf-file format from the supplementary section of: (Poongavanam et al., 2012).

Molecular Substructure mining using the MoSS algorithm

Molecular substructure searching (MoSS) was done using Borgelt`s MoSS implementation integrated into the KNIME software package. MoSS was published in 2002 and is explained in detail therein (Borgelt, Berthold, & Blvd, 2002; Borgelt, Berthold, & Patterson, 2005;

Hofer, Borgelt, & Berthold, 2004). Briefly, MoSS is an algorithm that finds discriminative substructures/fragments in data sets. These fragments occur more often as substructures of active molecules and at the same time occur rarely in the set of inactive molecules. Therefore, such fragments can be considered to be useful in discriminating the active from the inactive class. In order to identify such fragments MoSS adapts the algorithms of association rule induction - *Apriori* and *Eclat* (Agrawal, Imieliński, & Swami, 1993; Zaki & Parthasarathy, 1997) - and thereby represents molecules as attributed graphs and carries out a depth-first search on a tree of fragments (Borgelt et al., 2002). Going down one level in this search tree means extending a fragment by adding a bond or an atom to it (Borgelt et al., 2005). The MoSS algorithm allows for size-based pruning (the size of fragments can be adjusted by the user), support-based pruning (fragments that occur below a user-defined threshold are discarded) and structural pruning (for avoiding redundant searches). For the calculation of the most frequent substructures within the substrates and also the non-substrates the following parameters were used: minimum support in "focus"-class = 5%, minimum fragment size = 1, maximum fragment size = 100. The "focus" was the class of ABCB1-substrates in the search for fragments frequent in the "actives" and the class of ABCB1-non-substrates in the search for frequent "inactive" fragments. The support is the number of molecules that contain a certain fragment.

Structural Fingerprints

Structural fingerprints were calculated using the open-source RDKit-toolkit (http://www.rdkit.org/) implemented in the KNIME package (http://www.knime.org/). Seven different fingerprints were calculated. The calculated fingerprints include: (i) ECFP: circular fingerprint using the Morgan algorithm and connectivity invariants, (ii) FCFP: circular fingerprint using the Morgan algorithm and feature invariants, (iii) AP: atom-pair fingerprint, (iv) torsion: topological-torsion fingerprint, (v) RDKit: Daylight-like topological fingerprint, (vi) Avalon: fingerprint implemented in the RDKit from the Avalon toolkit (http://sourceforge.net/p/avalontoolkit/), (vii) layered: an substructure matching fingerprint. For circular fingerprints the radius for atomic environments to be considered was set to 2. The number of bits was set to 1024.

Calculation of MoSS-guided FP-SIBAR Descriptors

In order to calculate MoSS-guided FP-SIBAR descriptors the following workflow was employed: (i) calculation and identification of discriminatory fragments for the 243 ABCB1-substrates as well as for the 241 ABCB1 non-substrates, (ii) generation of a set of substrate molecules that consists of all substrates that contain at least one of the identified fragments and removal of duplicate entries (this yielded the set of potential active reference compounds), step (ii) was also carried out for the non-substrates (this yielded the set of potential inactive reference compounds), (iii) selection of 10 substrate reference molecules and 5 non-substrate reference molecules on basis of Willett`s maximum-dissimilarity algorithm (Willett, 1999). For substrate reference molecules digoxin was used as starting point and for the non-substrates NSC268251 was used as starting point for selecting the most diverse molecules. Digoxin was selected on basis of its potential role to elicit ABCB1-mediated drug-drug interactions (Giacomini et al., 2010). The 15 reference molecules were

removed from the 484 compound data set. The reason why 10 active and only 5 inactive molecules were selected was based on the fact that more discriminatory substructures were retrieved for the substrate class, (iv) calculation of the seven types of molecular fingerprints described above for the remaining 469 ABCB1 substrate/non-substrate molecules and the 15 reference molecules, (v) the fingerprints were used for calculating Tanimoto coefficients as similarity indices between the remaining 469 molecules under consideration and the 15 reference molecules. (vi) Finally, the MoSS-guided FP-SIBAR matrix was used as input for RF modelling. This workflow is graphically depicted in the following figure.

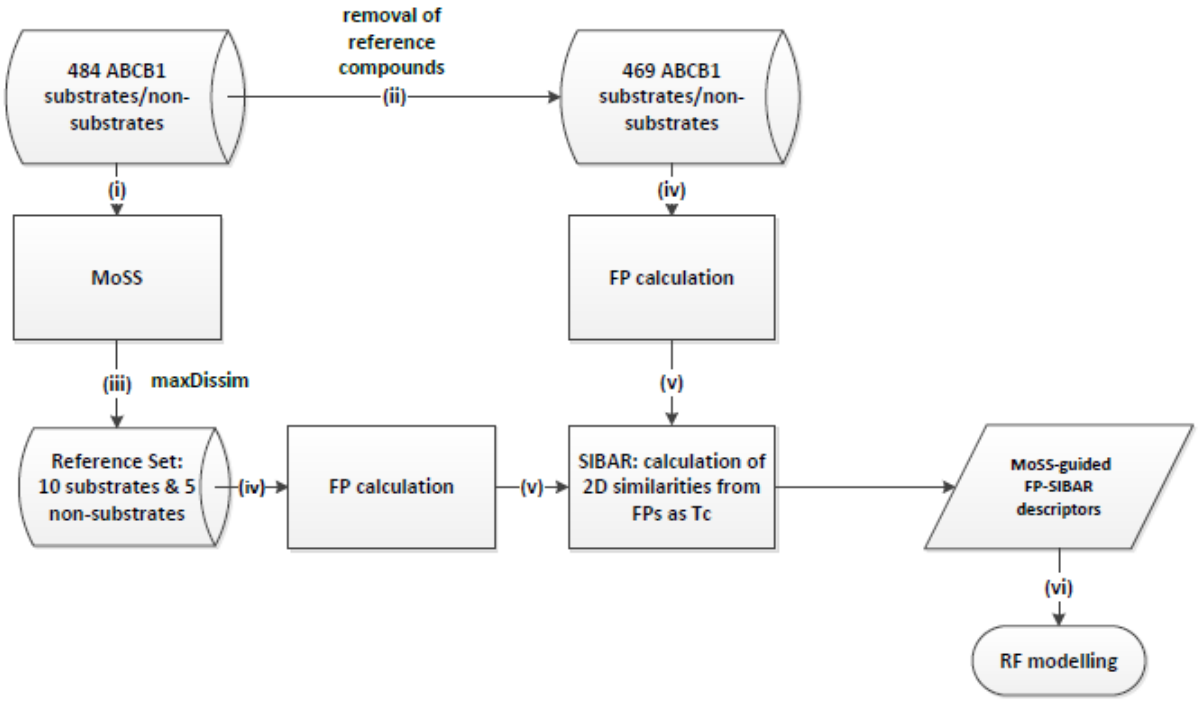


Figure 1: Workflow for the calculation of MoSS-guided SIBAR descriptors.

In summary, the data set of 484 (243 substrates/241 non-substrates) molecules was split into a training set of 469 (233 substrates/236 non-substrates) molecules and 15 references molecules, which were both submitted for subsequent similarity calculation, which finally returned the *469 x 15* MoSS-FP-SIBAR descriptor input matrix for consecutive random Forests modelling.

Comparison of MoSS-FP-SIBAR descriptors to "conventional" SIBAR descriptors

The MoSS-FP-SIBAR approach differs substantially from the "conventional" SIBAR approach (previously used by Klein et al., Kaiser et al., Zdrazil et al., Thai et al., and Schwaha et al. (Kaiser, Zdrazil, & Ecker, 2005; Klein et al., 2002; Schwaha & Ecker, 2009; Thai & Ecker, 2009; Zdrazil et al., 2007)) in three main aspects. First, the original implementation uses continuous values of structural/physicochemical descriptors, such as cLogP, HBA, whereas MoSS-FP-SIBAR uses binary fingerprints as input for similarity calculations. Second, discrete Tanimoto coefficients (Tc) in the range [0,1] are employed in MoSS-FP-SIBAR rather than the continuous Euclidean distance that was previously used. Third, Zdrazil et al. and Thai et al. pointed out that the choice of the reference set is crucial for the predictive

power of the SIBAR descriptors in QSAR modelling (Thai & Ecker, 2009; Zdrazil et al., 2007). Thai et al. showed that reference molecules selected from the set of known actives, i.e. reference molecules that are specifically tailored towards the biological problem performed best (Thai & Ecker, 2009). In this contribution, this concept is developed further. Instead of selecting reference molecules solely by taking into account the biological activity of the compounds under consideration, the reference set is also tailored towards the medicinal chemistry of the data set by also considering class discriminatory fragments. The common denominator of the conventional SIBAR approach and MoSS-FP-SIBAR is the use of similarity indices as input for subsequent modelling.

Classification Algorithm: Random Forests

In this contribution Breiman`s random Forests (RF) algorithm was used for classification of the data set consisting of 469 input molecules described by 15 MoSS-derived FP-SIBAR descriptors (Breiman, 1999). RF belongs to the class of ensemble algorithms, similar to boosting and bagging (Bruce, Melville, Pickett, & Hirst, 2007; Svetnik et al., 2003, 2005). An RF is comprised of a compilation of recursive decision trees which all give a vote to all training instances that is weighted and finally summed up to yield the final classification result. The main difference of RF compared to bagging is that it additionally employs random feature sampling – hence the name. For a detailed description of this algorithm the reader is referred to the original article (Breiman, 1999) or to Chapter 15 of the textbook of Hastie and Tibshirani (Hastie, Tibshirani, & Friedman, 2011). RF models were validated using 10-fold cross-validation.

Here, the "Tree Ensemble Learner" node implemented in the KNIME package was used with the following hyperparameters: split criterion: Gini Index; tree depth: unlimited; number of models = 500; instance sampling: 70%; attribute sampling: square root of features.

Software and miscellaneous calculations

Tanimoto coefficients, the most diverse reference molecules and the ROC analysis were computed in *R*. For the calculation of the similarity matrix containing Tc values the *fingerprint* package of Rajarshi Guha was used. The *fp.sim.matrix()* function was used here. In order to select the most diverse reference molecules for both classes the *maxDissim()* function in the *caret* package developed by Max Kuhn and obtained from CRAN was used. For computing model-independent variable importances the function *filterVarImp()* also contained in the *caret* was used.

Results and Discussion

Characterization of the Data Set by Molecular Substructures (MoSS)

The data set of 484 ABCB1 substrates/non-substrates was subjected to molecular substructure searching using Borgelt`s MoSS algorithm with the aim to identify class-specific fragments (Borgelt et al., 2005). For the ABCB1 substrates 59 discriminatory fragments were identified. The top ranked fragment showed a frequency of 20.1% within the class of active molecules and was identified as an integrated part of the chemical structure of 49 substrates. MoSS was also applied to the non-substrates. For this class 8 fragments were extracted. The top ranked fragment showed a frequency of 7.4% within the class of inactive molecules. It is interesting

to note that more fragments were identified for the substrate class and that the fragments obtained from the substrates show much higher frequency values compared to the frequency values of the non-substrates. A potential explanation for this might be that the class of non-substrates showed higher intra-class diversity and therefor a smaller number of frequently occurring substructures could be extracted from this class than from the obviously more homogeneous class of actives. The chemical formulas of the top ranked extracted fragments for ABCB1 substrates and non-substrates are represented in Figure 2 and Figure 3. Those active molecules that contained at least one of the 59 substructures were extracted into a separate data matrix and subsequently submitted to Willett`s maximum dissimilarity searching algorithm in order to extract the 10 most diverse molecules from the set of actives that contain at least one of the discriminatory fragments. The procedure was repeated for the non-substrates resulting in 5 diverse molecules. These 15 diverse molecules were selected in order to serve as reference molecules for the MoSS-FP-SIBAR descriptor calculation.

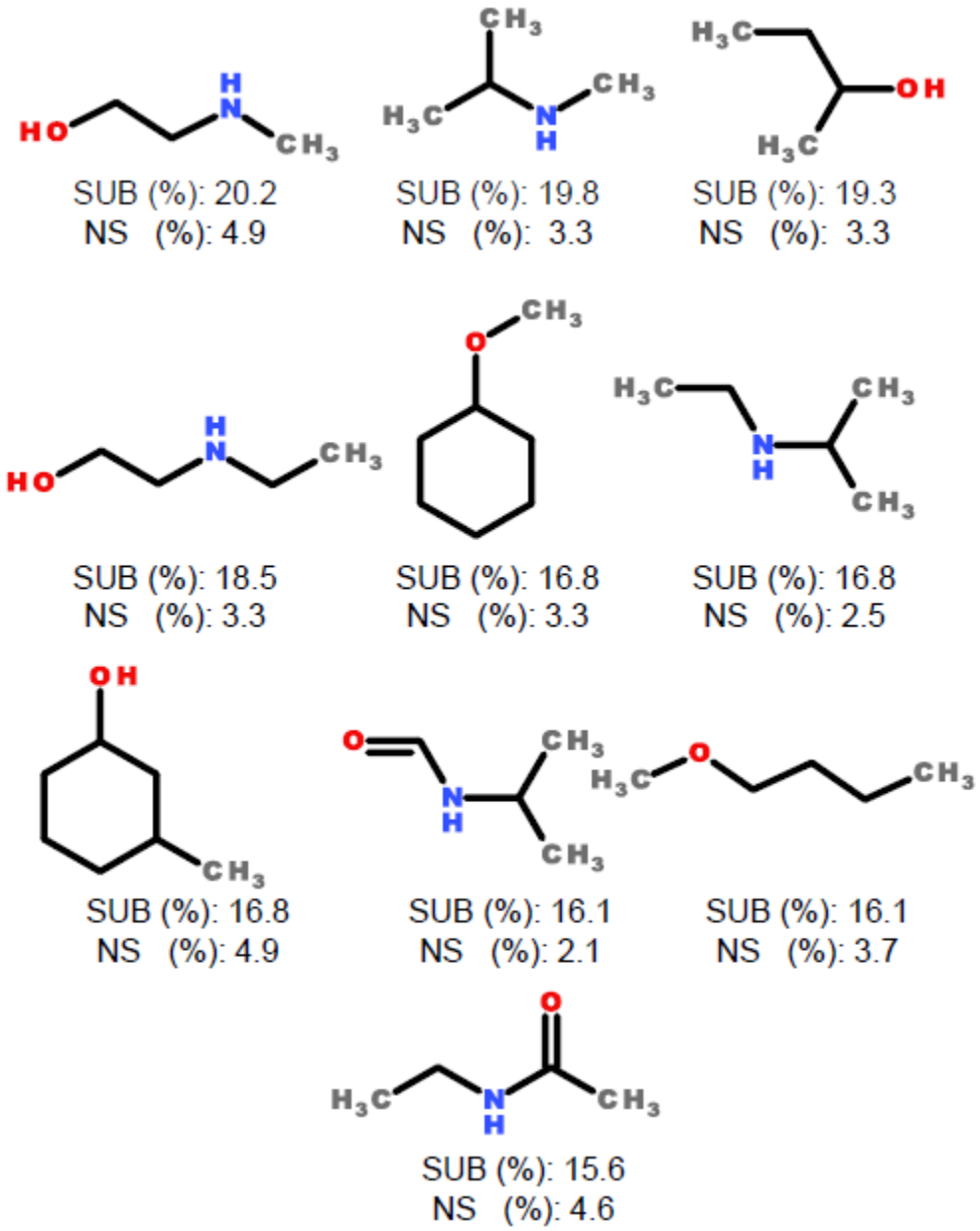


Figure 2: Structures of the top 10 molecular fragments frequently found in ABCB1-substrates.

Those active molecules that contained at least one of the 59 substructures were extracted into a separate data matrix and subsequently submitted to Willett`s maximum dissimilarity search algorithm in order to extract the 10 most diverse molecules from the set of actives that contain at least one of the discriminatory fragments. The procedure was repeated for the non-substrates resulting in 5 diverse molecules. The reason why we selected twice as many reference molecules for the active class is based on the fact that this class comprised much more fragments than the non-substrate class. These 15 diverse molecules were selected in order to serve as reference molecules for the MoSS-FP-SIBAR descriptor calculation.

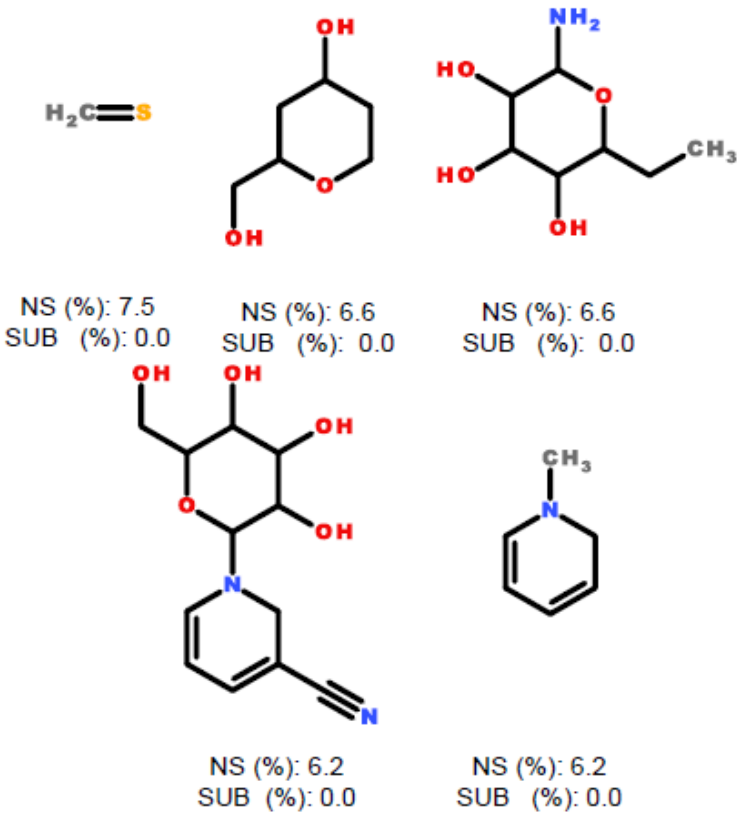


Figure 3: Structures of the top 5 molecular fragments frequently found in ABCB1-non-substrates

The chemical structures of the 10 reference substrates are shown in Figure 4 and the chemical structures for the 5 reference non-substrates are shown in Figure 5.

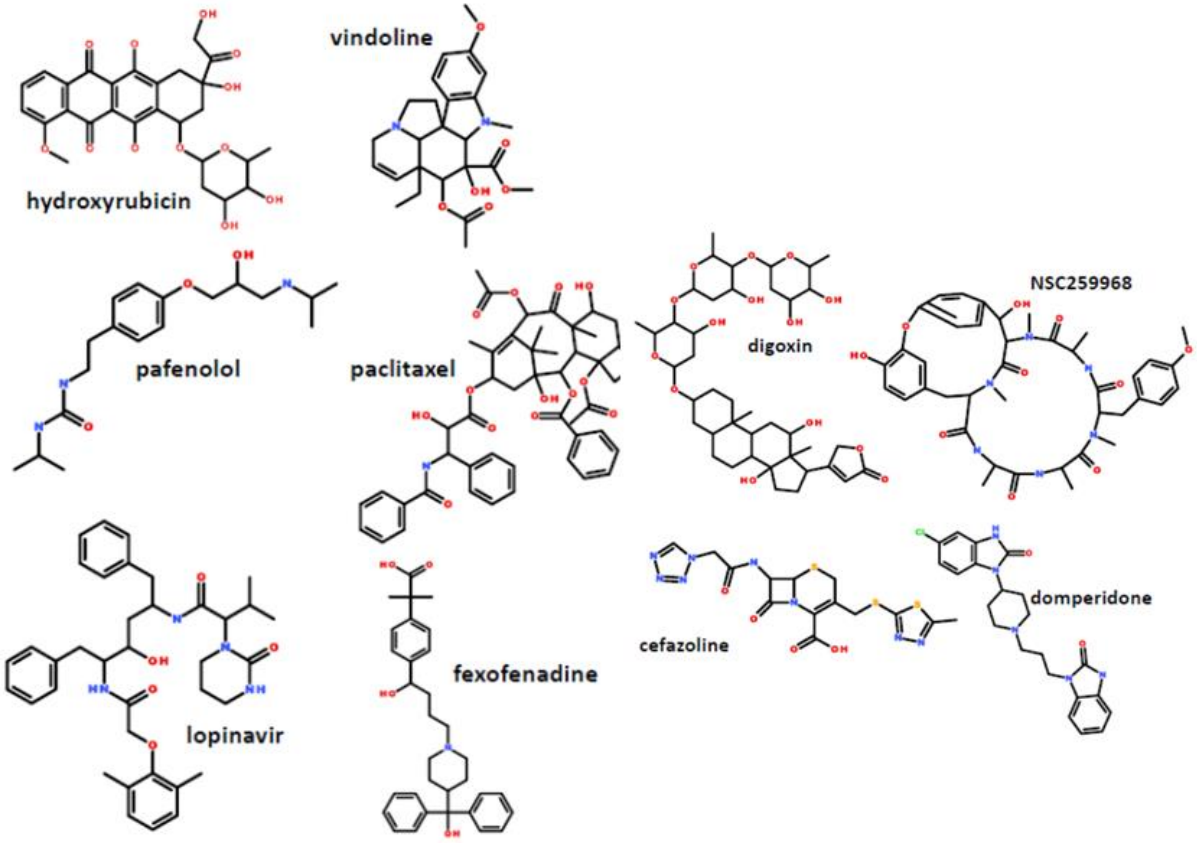


Figure 4: Chemical structures of the 10 ABCB1-substrates which serve as reference molecules

For the remaining 469 molecules which constituted the final training set for model evaluation as well as for the 15 reference molecules, seven different fingerprint types were calculated. Subsequently, the 15 reference molecules served as input for calculating Tanimoto coefficients as similarity values that substitute here for conventional descriptors.

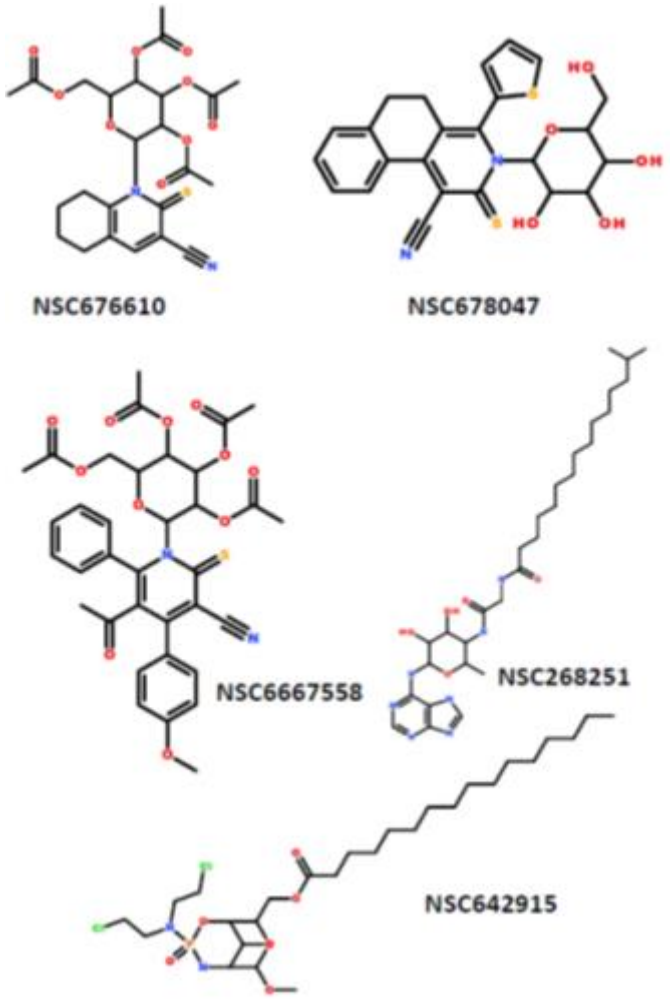


Figure 5: Chemical structures of the 5 ABCB1 non-substrates, which serve as reference molecules

Random Forests classification models using MoSS-FP-SIBAR descriptors

In order to evaluate the predictive power of the MoSS-FP-SIBAR descriptors we subjected the ABCB1 substrate/non-substrate data set to RF modelling. Additionally, we also constructed RF models using the mere fingerprints (FP-only) for the molecules for comparative purposes. Furthermore, another SIBAR set was built to serve for an additional comparison. This SIBAR set differs from the MoSS-FP-SIBAR set in that way that 10 active and 5 inactive molecules were randomly selected to serve as reference molecules instead of sampling them from molecules that contain discriminative fragments. This set was termed rand-SIBAR set. This procedure was repeated 10 times in order to account for random sampling error. In total two different comparisons were anticipated. First the comparison of MoSS-FP-SIBAR descriptors to standard fingerprints (FP-only) and second, the comparison of the MoSS-FP-SIBAR descriptors to SIBAR descriptors that contained reference molecules derived from random sampling (rand-SIBAR). The results for the ABCB1 substrate/non-substrate data set consisting of 469 molecules on basis of the different fingerprint are shown in Table 1. Results are reported as average from 10-fold cross-validation. Bold letters in Table 1 indicate the best performing models and italic letters indicate the models with the worst performance.

Table 1: Random Forests classification results for the different models using 10-fod CV. TP=true positives, TN=true negatives, FP=false positives, FN=false negatives, MCC= Matthew`s correlation coefficient; G-Mean = geometric mean encorporating sensitivity and specificity, Sens=sensitivity, Spec=specificity, Pr1=precision on substrates, Pr0=precision on non-substrates.

FP-only	TP	TN	FP	FN	MCC	Acc	G-Mean	Sens	Spec	Pr1	Pr0
ECFP	170	180	56	63	**0.49**	0.75	0.75	0.73	0.76	0.75	0.74
FCFP	162	190	46	71	**0.50**	0.75	0.75	0.70	0.81	0.78	0.73
AP	89	225	11	144	0.41	0.67	0.60	0.38	0.95	0.89	0.61
torsion	65	229	7	168	*0.35*	0.63	0.52	0.28	0.97	0.90	0.58
rdk	192	137	99	41	0.42	0.70	0.69	0.82	0.58	0.66	0.77
avalon	59	233	3	174	*0.36*	0.62	0.50	0.25	0.99	0.95	0.57
Layer	202	142	94	31	0.49	0.73	0.72	0.87	0.60	0.68	0.82
rand-SIBAR*											
ECFP	170	149	87	63	**0.36**	0.68	0.68	0.73	0.63	0.66	0.70
FCFP	130	191	45	103	**0.38**	0.68	0.67	0.56	0.81	0.74	0.65
AP	165	153	83	68	0.36	0.68	0.68	0.71	0.65	0.67	0.69
torsion	163	136	100	70	0.28	0.64	0.63	0.70	0.58	0.62	0.66
rdk	119	163	73	114	*0.20*	0.60	0.59	0.51	0.69	0.62	0.59
avalon	123	188	48	110	0.34	0.66	0.65	0.53	0.80	0.72	0.63
layer	109	181	55	124	*0.25*	0.62	0.60	0.47	0.77	0.66	0.59
MoSS-FP-SIBAR											
ECFP	180	159	77	53	**0.45**	0.72	0.72	0.77	0.67	0.70	0.75
FCFP	140	201	35	93	**0.47**	0.73	0.72	0.60	0.85	0.80	0.68
AP	175	163	73	58	0.44	0.72	0.72	0.75	0.69	0.71	0.74
torsion	173	146	90	60	0.36	0.68	0.68	0.74	0.62	0.66	0.71
rdk	129	173	63	104	*0.29*	0.64	0.64	0.55	0.73	0.67	0.62
avalon	133	198	38	100	0.43	0.71	0.69	0.57	0.84	0.78	0.66
layer	119	191	45	114	*0.34*	0.66	0.64	0.51	0.81	0.73	0.63

*rand-SIBAR: results report the mean of the 10 different rand-SIBAR sets.

The results of the RF classification models in Table 1 were evaluated the models on basis of

- Matthews correlation coefficient (MCC): ((TPxTN)-(FPxFN)/SQRT((TP+FP)x(TP+FN)x(TN+FP)x(TN+FN))),
- G-mean: (SQRT(Sens x Spec)) and
- total accuracy (Acc): ((TP+TN)/(TP+FP+FN+TN)),
- additionally sensitivity (Sens), specificity (Spec), precision on actives (Pr1), precision on inactives (Pr0) – for formulas see (Demel, Janecek, Gansterer, & Ecker, 2009; Demel, Janecek, Thai, Ecker, & Gansterer, 2008; Demel, Kraemer, Ettmayer, Haaksma, & Ecker, 2010; Hillebrecht & Klebe, 2008)

These performance measures are widely used in the field of machine learning (Hastie et al., 2011). Most of these measures provide values in the range of [0,1]. The only exception is MCC which reports performance in the range of [-1,1]. All these measures can be interpreted

in the same way: the closer the measure to 1, the better the model. Regarding MCC, a value > 0.4 is considered to reflect a predictive model . From Table 1 one can compare the MoSS-FP-SIBAR descriptor models to the FP-only models and the rand-SIBAR models. It has to be noted that the rand-SIBAR results are reported as mean of the 10-fold cross-validation of the 10 different randomly selected rand-SIBAR sets. It can be seen that the MoSS-FP-SIBAR descriptors when calculated on basis of circular fingerprints, i.e. ECFP and FCFP show a similar performance (MCC (ECFP): 0.45, MCC (FCFP): 0.47) than the model constructed on these circular fingerprints only (MCC (ECFP): 0.49, MCC (FCFP): 0.50). This performance is also comparable to the published fingerprint-based model of Poongavanam et al. who reported RF performances of 0.49 for MCC (Poongavanam et al., 2012). Furthermore, it is shown that ECFP and FCFP fingerprints outperform other types of fingerprints. The finding that circular fingerprints give the best performance is not surprising since these type of fingerprints was shown to be superior to other 2D fingerprints and are also capable to outperform even 3D Phase shape fingerprints (Hu et al., 2012). Nevertheless, the performance differences observed for other fingerprint-types further support previous findings that the choice of the feature space, which forms the basis of similarity calculations, is essential, for the performance of SIBAR descriptors (Schwaha & Ecker, 2009; Thai & Ecker, 2009). Both the MoSS-FP-SIBAR descriptor model as well as the FP-only model outperform the rand-SIBAR model. The finding that MoSS-FP-SIBAR outperforms rand-SIBAR supports the rationale for selecting reference molecules on basis of discriminating substructures which is proposed here. However, considering the fact that MoSS-FP-SIBAR is not superior to the FP-only model, immediately raises the question if the observed equal performance between the two methods outweigh the complexity of the MoSS-FP-SIBAR approach. In order to answer this question, one is reminded that the FP-only model is built on a much higher dimensional input space (formally 1024 bits), than the MoSS-FP-SIBAR model, which is only constructed out of 15 Tanimoto coefficient vectors. Considering now the fact that both methods return highly comparable performance measures, shows that the low-dimensional MoSS-FP-SIBAR descriptors convey a comparable amount of chemical information (without a loss in predictive performance) than the high dimensional fingerprints. From this viewpoint MoSS-FP-SIBAR is shown to be a putative useful method for dimensionality reduction and could be regarded as an alternative for principal component calculation or related methods. In summary, it is confirmed on basis of RF modelling and employing 10-fold cross-validation that the MoSS-FP-SIBAR descriptors are capable of providing predictive models.

Interpretation of the Moss-FP-SIBAR Descriptors

As mentioned above, the best MoSS-FP-SIBAR descriptor RF model on basis of ECFP fingerprints is only constructed out of 15 input variables, namely the 15 reference molecules. Taking into account that the input variables represent Tc values for the molecules in the data set to the reference molecules, suggests that these descriptors might also be valuable for model interpretation purposes. Therefore, RFs intrinsic variable importance measure, which assesses the mean decrease in classification accuracy (MDA) was applied to identify the reference molecules that contributed most to the model. The details on the calculation of MDA are in explained in Breiman et al. and exemplary applications of this measure are provided in Svetnik et al. (Breiman, 1999; Svetnik et al., 2003). Figure 6 below visualizes the

ranking of the 15 reference molecules (variables) according to their importance in the RF model. It is interesting to note that the top-ranked variables denote only refence molecules belonging to the substrate class, with digoxin and domperidone being the top-ranked reference molecules (variables) according to MDA.

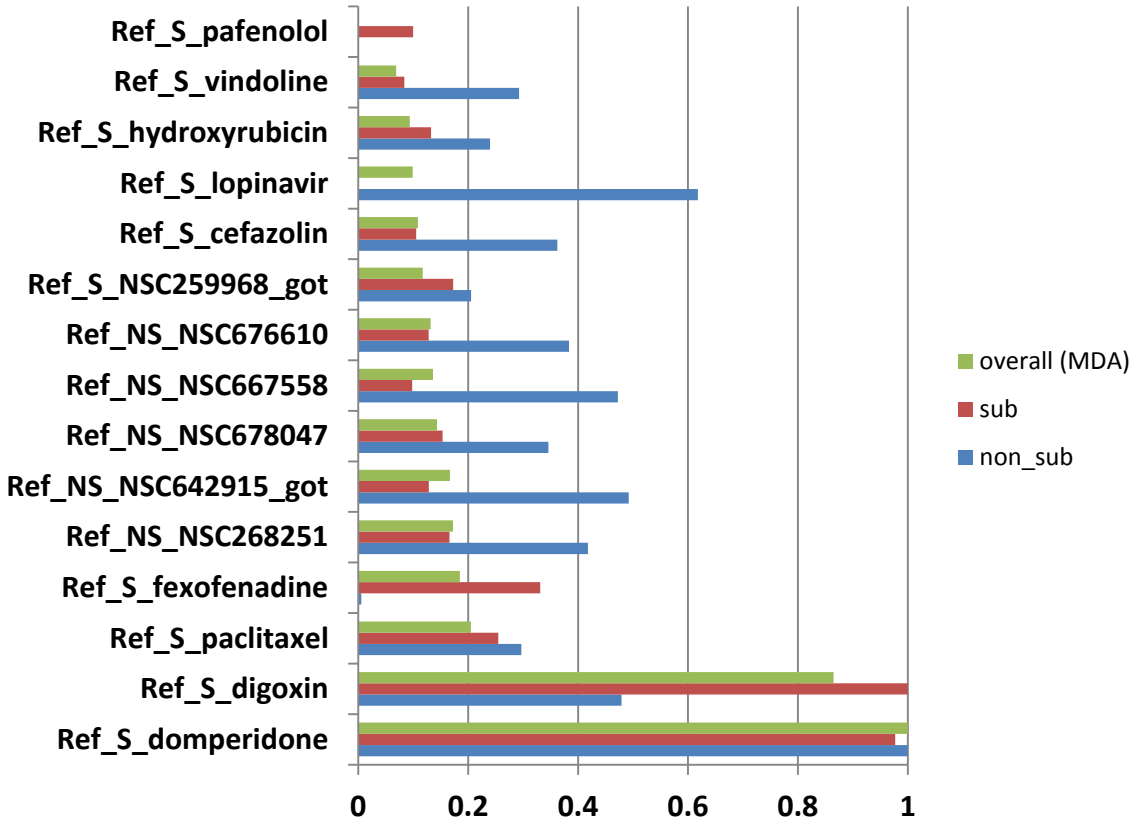


Figure 6: RF variable importance for the MoSS-FP-SIBAR model on basis of circular ECFP-fingerprints. Relative importances are shown. Variables are ordered with respect to decreasing overall performance, reported as mean decrease in classification performance (MDA).

The fact that digoxin is showing the highest predictive contribution to the model is of particular interest. First of all digoxin is a well-known substrate of ABCB1 and also can give rise to mediate serious adverse drug reactions when co-administered with other molecules known to interact with ABCB1 in the clinic (Ding et al., 2004). Additionally, Giacomini et al. proposed that digoxin is a suitable probe substrate for the *in-vivo* assessment of the ABCB1 substrate liability of new molecular entities (Giacomini et al., 2010). Next, we wanted to compare the model-specific variable importance measure MDA incorporated into the RF algorithm to another model independent measure. This comparison was done via model-independent Receiver Operating Characteristic (ROC) analysis. Therefore, all the 15 Tc-vectors that constitute the MoSS-FP-SIBAR descriptors were used individually to construct ROC curves. ROC analysis was conducted for each reference molecule separately. Since the similarity measure Tc is in the range of [0,1], it is directly useable for computing a series of cutoffs to predict each class (in a model independent fashion). The sensitivity and specificity for each of these cutoff points was calculated. For deriving the area under the ROC curve the trapezoidal rule was used. The results are graphically visualized in the ROC plot depicted in

Figure 7A (more characteristics on the individual model-independent predictive potential of the 15 reference molecules can be found in the supplementary information section accompagnying this contribution).

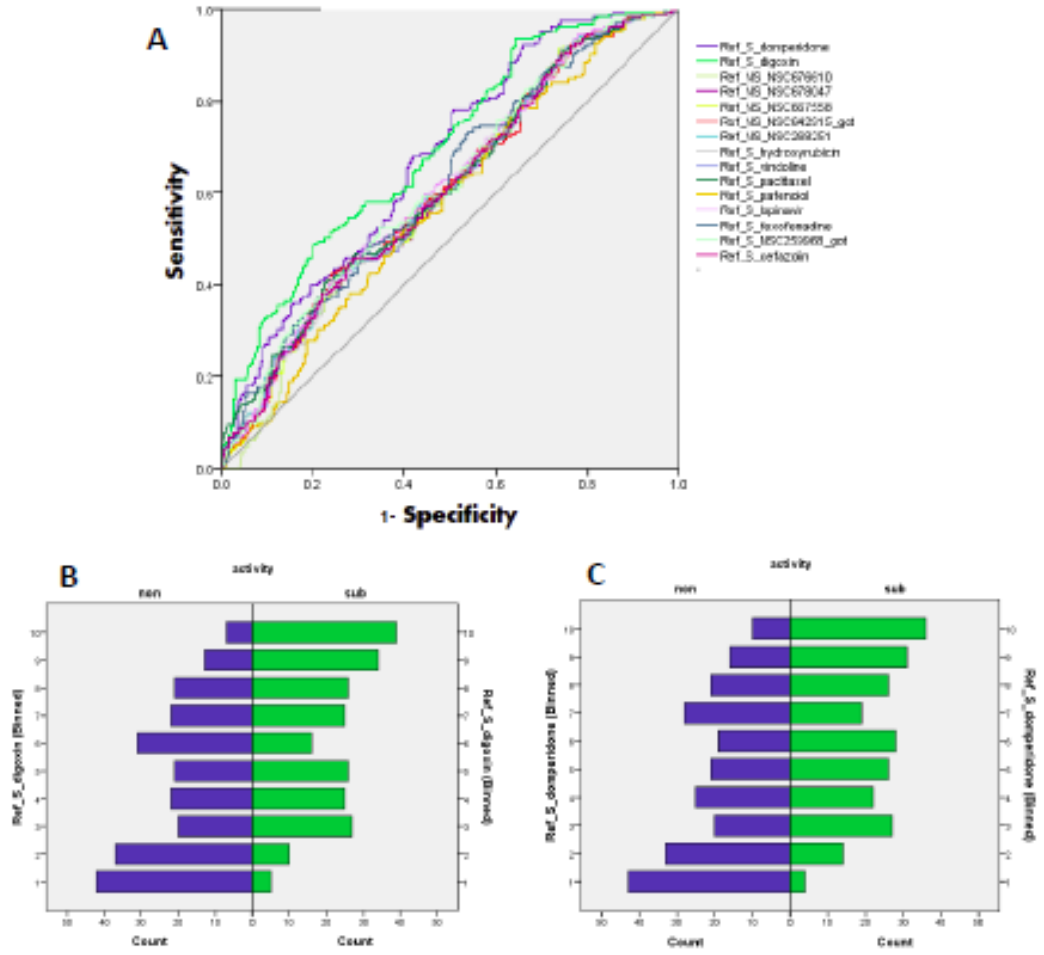


Figure 7A-C: Model-independent evaluation of the class separating properties of the reference used for MoSS-FP-SIBAR descriptors on basis of ECFP fingerprints. A) ROC curves displaying the model-independent properties of all 15 reference molecules. B) and C) show the class-specific distribution of the counts of the binned Tc-vector for B) digoxin and C) domperidone. For binning the Tc-vectors were separated into 10 bins of equal width. Blue bars in B) and C) depict the counts of the non-substrates whereas green bars in B) and C) depict the counts for the substrates.

Indeed, the ROC analysis in the figure above further supports the predictive power of digoxin (light green) and domperidone (pink). Both reference molecules exhibit an area under the ROC curve higher than 0.65 with small standard errors and small confidence intervals (see also supplementary information). Furthermore, the plots of the class-specific distribution of the binned Tc values for digoxin and domperidone support their predictive capability for this data set. To further get a better understanding of the importance of these two reference molecules, the RF model, which consisted of 500 trees was manually searched for trees that resemble these two reference molecules. One example of a decision tree found that is on the one hand simple and utilizes these two reference molecules only is shown in Figure 8.

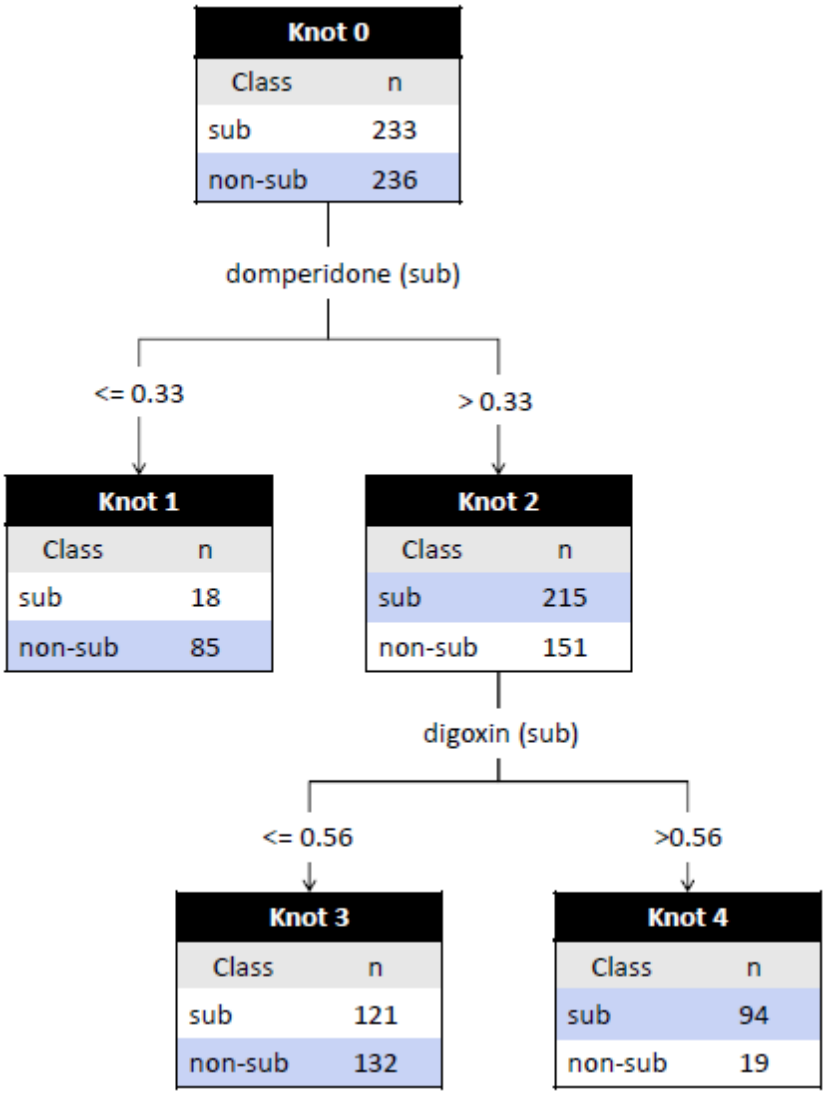


Figure 8: Example of one decision tree (out of the 500 trees of the best RF model) that utilizes the digoxin and domperidone as reference molecules.

The decision tree that is shown above serves the purpose to provide an illustrative example for the interpretation of MoSS-FP-SIBAR descriptors. It is interpreted as follows: molecules that show a similarity to domperidone smaller or equal than Tc=0.33 are likely to be non-substrates. Among those molecules that are more related to domperidone (Tc>0.33) are further separated with respect to their similarity to digoxin. The remaining molecules that are more similar (Tc>0.55) to digoxin are also likely to be substrates, whereas those molecules unrelated to this reference molecules (Tc<=0.55) are considered to resemble non-substrates. It is important to note, that this tree represent only one tree out of the 500 trees that constitute the full ensemble and that this tree is not validated by any means (performance characteristics can be found in the supplementary information section). Nevertheless, it demonstrates impressively that MoSS-FP-SIBAR descriptors can be easily interpreted and encode concepts that are familiar to medicinal chemists.

Conclusion

In this study a novel approach to derive SIBAR descriptors, MoSS-FP-SIBAR, is preliminarily outlined and critically appraised. The approach utilizes fingerprints rather than physicochemical descriptors and applies the discrete Tc rather than the continuous Euclidean distance measure compared to the "conventional" SIBAR approach. Furthermore, the reference set is not only tailored towards the pharmacological problem, but also incorporates the medicinal chemistry of the data set under consideration into the selection process. In order to get a preliminary idea of the feasibility of MoSS-FP-SIBAR descriptors, this approach was directly compared to a simple fingerprint-based model and SIBAR set, that was based on random sampling for reference set generation. This comparison revealed that on principle MoSS-FP-SIBAR shows a similar performance to a fingerprint-based model. At first glance, this seems to be disappointing if one takes into account that the calculation of these descriptors also embodies much more steps and is therefore more labour-intensive than the calculation of mere fingerprints. However, the primary advantage of this approach is that it also incorporates an alternative dimensionality reduction method. The high dimensionality of the fingerprint is reduced to a small set of reference molecules. Furthermore, MoSS-FP-SIBAR descriptors encode Tc-vectors that can be easily interpreted and eventually might be directly related into chemical optimization processes. Nevertheless, this approach should further be evaluated using different classification problems and more rigorous validation methods in future work. It is hoped that the MoSS-FP-SIBAR descriptors might receive a lot of acceptance from the community, since it encodes principles that are familiar to the medicinal chemist and can be easily incorporated into future drug design projects.

References

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93* (pp. 207–216). New York, New York, USA: ACM Press. doi:10.1145/170035.170072

Blundell, T. L., Jhoti, H., & Abell, C. (2002). High-throughput crystallography for lead discovery in drug design. *Nature Reviews Drug Discovery*, *1*(1), 45–54. doi:10.1038/nrd706

Borgelt, C., Berthold, M. R., & Blvd, G. (2002). Mining Molecular Fragments : Finding Relevant Substructures of Molecules. *IEEE International Conference on Data Mining (ICDM 2002, Maebashi, Japan)* (pp. 51–58).

Borgelt, C., Berthold, M. R., & Patterson, D. E. (2005). Molecular Fragment Mining for Drug Discovery. *Workshop Open Source Data Mining Software (OSDM'05, Chicago, IL)* (pp. 6–15).

Breiman, L. (1999). Random Forestss, 1–35.

Browne, W. J., North, A. C., & Phillips, D. C. (1969). A possible three-dimensional structure of bovine-lactalbumin based on that of hen`s egg-white lysozyme. *J. Mol. Biol.*, *42*, 65–86.

Bruce, C. L., Melville, J. L., Pickett, S. D., & Hirst, J. D. (2007). Contemporary QSAR classifiers compared. *Journal of Chemical Information and Modeling*, *47*(1), 786–99. doi:10.1021/ci600332j

Chen, Z., Li, H., Zhang, Q., Bao, X., Yu, K., Luo, X., Zhu, W., et al. (2009). Pharmacophore-based virtual screening versus docking-based virtual screening: a benchmark comparison against eight targets. *Acta Pharmacologica Sinica*, *30*(12), 1694–708. doi:10.1038/aps.2009.159

Demel, M. A., Janecek, A. G. K., Gansterer, W. N., & Ecker, G. F. (2009). Comparison of Contemporary Feature Selection Algorithms: Application to the Classification of ABC-Transporter Substrates. *QSAR & Combinatorial Science*, *28*(10), 1087–1091. doi:10.1002/qsar.200860191

Demel, M. A., Janecek, A. G. K., Thai, K.-M., Ecker, G. F., & Gansterer, W. N. (2008). Predictive QSAR Models for Polyspecific Drug Targets: The Importance of Feature Selection. *Current Computer - Aided Drug Design*, *4*(2), 91–110. doi:10.2174/157340908784533256

Demel, M. A., Kraemer, O., Ettmayer, P., Haaksma, E., & Ecker, G. F. (2010). Ensemble Rule-Based Classification of Substrates of the Human ABC-Transporter ABCB1 Using Simple Physicochemical Descriptors. *Molecular Informatics*, *29*(3), 233–242. doi:10.1002/minf.200900079

Ding, R., Tayrouz, Y., Riedel, K.-D., Burhenne, J., Weiss, J., & Mikus, G. (2004). Substantial pharmacokinetic interaction between digoxin and ritonavir in healthy volunteers. *Clin. Pharmacol. Ther.*, *76*(1), 73–84.

Giacomini, K. M., Huang, S.-M., Tweedie, D. J., Benet, L. Z., Brouwer, K. L. R., Chu, X., Dahlin, A., et al. (2010). Membrane transporters in drug development. *Nature Reviews Drug Discovery*, *9*(3), 215–36. doi:10.1038/nrd3028

Greer, J. (1980). Model for haptoblobin heavy chain based upon structural homology. *Proc Natl Acad Sci U S A*, *77*, 3393–3397.

Hastie, T., Tibshirani, R., & Friedman, J. (2011). *The Elements of Statistical Learning* (p. 711).

Hillebrecht, A., & Klebe, G. (2008). Use of 3D QSAR models for database screening: a feasibility study. *Journal of Chemical Information and Modeling*, *48*(2), 384–96. doi:10.1021/ci7002945

Hofer, H., Borgelt, C., & Berthold, M. R. (2004). Large Scale Mining of Molecular Fragments with Wildcards. *Inteligent Data Analysis*, *8*, 495.504.

Hu, G., Kuang, G., Xiao, W., Li, W., Liu, G., & Tang, Y. (2012). Performance evaluation of 2D fingerprint and 3D shape similarity methods in virtual screening. *Journal of Chemical Information and Modeling*, *52*(5), 1103–13. doi:10.1021/ci300030u

Kaiser, D., Zdrazil, B., & Ecker, G. F. (2005). Similarity-based descriptors (SIBAR)--a tool for safe exchange of chemical information? *Journal of Computer-Aided Molecular Design*, *19*(9-10), 687–92. doi:10.1007/s10822-005-9000-8

Klein, C., Kaiser, D., Kopp, S., Chiba, P., & Ecker, G. F. (2002). Similarity based SAR (SIBAR) as tool for early ADME profiling. *Journal of Computer-Aided Molecular Design*, *16*(11), 785–93. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12825790

Kubinyi, H. (1998). Molekulare Ähnlichkeit. 1. Chemische Struktur und biologische Wirkung. *Pharmazie Unserer Zeit*, 92–106.

Martin, Y. C., Kofron, J. L., & Traphagen, L. M. (2002). Do structurally similar molecules have similar biological activity? *Journal of Medicinal Chemistry*, *45*(19), 4350–8. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12213076

Perutz, M. F., Kendrew, J. C., & Watson, H. C. (1965). Structure and function of hemoglobin. *J. Mol. Biol.*, *13*, 669–678.

Poongavanam, V., Haider, N., & Ecker, G. F. (2012). Fingerprint-based in silico models for the prediction of P-glycoprotein substrates and inhibitors. *Bioorganic & Medicinal Chemistry*. doi:10.1016/j.bmc.2012.03.045

Schwaha, R., & Ecker, G. F. (2008). The Similarity Principle – New Trends and Applications in Ligand-Based Drug Discovery and ADMET Profiling. *Scientia Pharmaceutica*, *76*(1), 5–18. doi:10.3797/scipharm.0802-05

Schwaha, R., & Ecker, G. F. (2009). Similarity Based Descriptors - Useful for Classification of Substrates of the Human Multidrug Transporter P-Glycoprotein? *QSAR & Combinatorial Science*, *28*(8), 834–839. doi:10.1002/qsar.200960051

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random Forests: a classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences*, *43*(6), 786–99. doi:10.1021/ci034160g

Svetnik, V., Wang, T., Tong, C., Liaw, A., Sheridan, R. P., & Song, Q. (2005). Boosting: an ensemble learning tool for compound classification and QSAR modeling. *Journal of Chemical Information and Modeling*, *45*(3), 786–99. doi:10.1021/ci0500379

Thai, K.-M., & Ecker, G. F. (2009). Similarity-based SIBAR descriptors for classification of chemically diverse hERG blockers. *Molecular Diversity*, *13*(3), 321–36. doi:10.1007/s11030-009-9117-0

Willett, P. (1999). Dissimilarity-based algorithms for selecting structurally diverse sets of compounds. *Journal of Computational Biology*, *6*(3-4), 447–57. doi:10.1089/106652799318382

Zaki, M. J., & Parthasarathy, S. (1997). New Algorithms for Fast Discovery of Association Rules. *Proc. 3rd Int. Conf. on Knowledge Discovery and Data Mining (KDD'97)* (pp. 283–296).

Zdrazil, B., Kaiser, D., Kopp, S., Chiba, P., & Ecker, G. F. (2007). 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. doi:10.1002/qsar.200610149

Author contribution

I planned the conceptual design of this contribution and carried out all the calculations. I am the sole author of this manuscript. The data set was retrieved from the indicated reference (see main text).

Supplementary Information

ST1: Characteristics of the ROC analysis

			asymptotic 95% confidence interval	
variables	AUC	standard error[a]	lower bound	upper bound
Ref_S_domperidone	**.675**	**.024**	**.628**	**.723**
Ref_S_digoxin	**.697**	**.024**	**.650**	**.744**
Ref_NS_NSC676610	.596	.026	.544	.647
Ref_NS_NSC678047	.603	.026	.552	.654
Ref_NS_NSC667558	.603	.026	.552	.654
Ref_NS_NSC642915_got	.607	.026	.556	.657
Ref_NS_NSC268251	.609	.026	.558	.659
Ref_S_hydroxyrubicin	.607	.026	.556	.657
Ref_S_vindoline	.605	.026	.554	.656
Ref_S_paclitaxel	.613	.026	.563	.664
Ref_S_pafenolol	.578	.026	.526	.629
Ref_S_lopinavir	.615	.026	.564	.665
Ref_S_fexofenadine	.618	.026	.567	.668
Ref_S_NSC259968_got	.618	.026	.568	.668
Ref_S_cefazolin	.607	.026	.556	.657

a. non-parametric assumption

ST2: Performance characteristic of the example decision tree utilising digoxin and domperidone:

	sub	non-sub
sub	**94**	**139**
non-sub	**19**	**217**
Acc	**Sens**	**Spec**
0.66	0.40	0.92

Chapter 15: Manuscript III - In-silico Machine Learning Models for the Characterization of Molecules Selectively Killing ABCB1-Overexpressing Multidrug-Resistant Cancer Cells

Abstract

Chemotherapy is currently one of the most effective treatment modalities for metastatic cancers. Over-expression of ABCB1 (P-glycoprotein, P-gp, MDR1), a member of the human, multidrug ATP-binding cassette (ABC)-transporter family, represents one of the major reasons for the development of a multidrug resistant (MDR) phenotype and is often associated with poor survival and severely compromises therapeutic success. One rather new principle to resolve MDR is the rational design of compounds that selectively kill cancer cells over-expressing this efflux transporter. Thus, predictive in-silico models provide versatile tools to characterize properties associated with MDR-selective agents. In this study we developed machine learning models for the classification of MDR-selective agents using three different algorithms, namely Random Forests, a Gradient Boosting Machine and a Support Vector Machine on a data set comprised of 86 compounds described by 30 molecular descriptors that have been selected using an unsupervised feature selection algorithm. The results provided in this study show that Random Forests clearly outperforms the other two methods on an external validation set with an overall correct classification rate of 83%. Furthermore it is shown that 2D-autocorrelation descriptors weighted by molecular mass represent in general the most useful features to characterize favorable and unfavorable properties of MDR-selective agents. These results represent good starting points that can be rationally used by experimental chemists to further expand the emerging field of MDR-targeted therapies as a new paradigm to overcome ABCB1-mediated cancer MDR.

Introduction

Cancer represents one of the major global health burdens mankind has to face nowadays. In 2008, about 12.7 million cancer cases and 7.6 million cancer deaths occurred worldwide. Although chemotherapy is currently one of the most effective ways to treat metastatic cancers, the occurrence of multi-drug resistance (MDR) has become a major impediment to cancer therapy and has been attributed the failure for treatment in over 90% of patients suffering from metastatic cancer resulting in relapse and eventually deaths of these patients [1]. Despite the fact, that MDR can arise from several, multiple causes, the over-expression of the ATP-binding-cassette (ABC) transporter ABCB1 (MDR1, P-glycoprotein, P-gp, CD243) has been associated as the primary mechanism for reducing intracellular drug concentrations and thereby compromising treatment success [2]. ABCB1 is a multidrug transporter that actively extrudes its substrates out of cancer cells. It represents a transmembrane protein being constituted of four essential domains, two α-helical transmembrane domains (TMDs) and two intracellular localized nucleotide-binding domains (NBDs) [3]. The TMDs are responsible for positioning in the cellular membrane and also for substrate recognition and binding, whereas the NBDs bind and hydrolyze ATP to provide energy for substrate transport [4, 5]. The hallmark of ABCB1 is its unusually broad polyspecificity [6]. It recognizes and transports hundreds of mainly large, hydrophobic and positively charged compounds that can have strikingly dissimilar chemical structures and pharmacological profiles, including clinically relevant drugs such as anticancer drugs, antiretroviral agents, immunosuppressive drugs and antiepileptics [2, 7-10]. The mostly hydrophobic substrates tend to partition into the lipid bilayer [11]. Thus, ABCB1 has been compared to a molecular "vacuum cleaner", which pulls substrates out of the cellular membrane before they can reach their intracellular target, thereby promoting MDR [12]. On grounds of the ability of ABCB1 to mediate a MDR phenotype to cancer cells, it was one of the first concepts to rationally design inhibitors of this

transporter [13-16]. Although, most in-vitro results with several generations of ABCB1 inhibitors were quite encouraging, successful modulation of MDR through blocking this efflux pump has not been successful in a clinical setting so far [2, 17]. Therefore, new strategies are urgently required to overcome ABCB1-mediated MDR

A very promising strategy to resolve MDR is to turn the mechanism of drug resistance into a weakness. Compounds inducing preferential cytotoxicity to ABCB1 expressing cancer cells have been recently described in the literature as "collateral sensitive" compounds [18]. Such molecules confer a type of "synthetic lethality" in which they exploit the genetic alteration (here ABCB1 overexpression) that confers drug resistance to other drugs [19]. In other words, cytotoxicity of a drug is enhanced by the presence of ABCB1, rather than reduced. Collateral sensitive compounds are also termed "MDR-selective" compounds in the context of ABCB1-overexpressing cancer cells. MDR-selectivity is easily determined in vitro by assessing the ratio of cytotoxicity of a compound in a MDR cell line and its parental cell line [19]. Molecules displaying a selectivity ratio >1 are classified as MDR selective agents [20, 21]. In 2009, the group of Gergely Szakacs screened the Developmental Therapeutics Program (DTP) drug activity set of 42,000 molecules for MDR-selective agents [21]. This search resulted in a data set containing the selectivity ratios mentioned above. Although, the detailed molecular mechanism of action is currently not well understood, it might be possible that the target for MDR-selective agents is not ABCB1 per se [22-24]. Nevertheless, the biochemical characterization of the MDR-selective agents showed that functional ABCB1 is required to confer the MDR-selective phenotype and additionally it was observed that ABCB1 expression decreased in cancer cells upon extended exposure [25]. This phenomenon also has been shown to be cell line independent. For a very promising thiosemicarbazone NSC73306 of the NCI60 drug sensitivity screen, it was additionally shown that cells are hypersensitive to this agent in proportion to ABCB1 function [26]. Furthermore, this MDR selectivity is abrogated by functional inhibition of ABCB1 or si-RNA mediated knock-down [27]. For this preclinical candidate, NSC73306, it was also proven that it is neither a substrate for ABCB1 nor an inhibitor. However, it was found that it serves as substrate for another ABC-transporter, namely ABCG2 [28].

In-silico exploration and characterization of data sets is an integral part of every successful drug development programme nowadays. Especially, the application of machine learning (ML) has gained more and more popularity during the past decades in various areas of drug design [29-34]. A ML QSAR model is usually computed by using an experimental data set that is constituted by a number or molecular descriptors (X) that encode the structural information of the compounds in a numerical manner and the corresponding biological activities (Y) for each molecule. The Y-variable can be categorical in case of a classification model or can be continuous in case of a regression model. A large variety of chemical descriptors are available that encode chemicals at different levels of complexity and sophistication. Additionally, a broad array of ML algorithms are available that differ remarkably with respect to classification performance, ability to model non-linear relationships and competence to provide interpretability of results. Hence, it is the elaborate combination of selected features in conjunction with a suitable modeling algorithm that can provide both; good classification performance and interpretability.

ABCB1 has been the center of many *in-silico* investigations that were mainly concentrating on identifying ABCB1-substrates [35-39] and ABCB1-inhibitors [40-42]. Summarizing these studies, it can be generalized that molecules interacting with ABCB1 are mainly characterized by higher molecular weight, are more lipophilic, are more flexible and often contain a nitrogen atom that seems to be involved in mediating cation-π interactions with aromatic amino acids compared to their

respective counterparts [35, 43, 44]. These results mirror the extensive efforts that have been made to characterize the ligands interacting with this transporter. However there is only limited SAR data available that characterizes MDR-selective agents [20, 21, 26]. One pharmacophore hypothesis highlights the importance of an electron-rich substitution at the N4 position of a series of isatin-ß-thiosemicarbazones that are derivatives of the preclinical candidate NSC73306 [20]. A QSAR analysis on derivatives of NSC73306 and a series of hydrazinylquinoline was based on highly complex WHIM-, GETAWAY and autocorrelation descriptors, but nevertheless returned models with excellent predictive power [21]. To this end, it is the primary objective of this study to provide the first ML models for classification of MDR selective and non-MDR selective compounds and to explore chemical features that are associated with the hypersensitive properties these compounds confer to ABCB1 over-expressing cancer cells. We made use of a data set that is comprised of 86 compounds which are categorized into MDR-selective and non-MDR-selective molecules. Structures were represented by a comprehensive set of 2D-dimensional molecular descriptors. In order to reduce the feature space with the focus to eliminate redundant and highly correlated descriptors and in an effort to maximize classification speed, unsupervised feature selection was applied. Using a Random Forests (RF) algorithm, a gradient boosting machine (GBM) and a support vector machine (SVM), several models were generated for the identification of MDR-selective agents on the reduced feature subset. Parameter tuning and model selection was carried out using three independent repetitions of 10-fold cross-validation. Finally the best performing models were validated on basis of an external test set. In order to gain insights into the molecular properties associated with the medicinal chemistry of MDR-selective compounds, measures of feature importance and partial dependence plots were employed for the model which performed best on the external test set.

Material and Methods

Data Sets

The chemical and pharmacological data employed in this study were taken from Türk et al.[21]. The pharmacological details are explained in detail therein. Briefly, Türk and colleagues assayed a broad collection of chemicals using a MTT cytotoxicity assay on the KB-V1 ABCB1 expressing cell line and the parental KB-3-1 cell line. The parental cell line does not express the ABC-transporter ABCB1. To determine the MDR selectivity of the compounds, they calculated the ratio of a compound`s IC50 against KB-3-1 cells divided by its IC50 against KB-V1 cells. A selectivity index >1 indicates that a compound kills ABCB1-expressing cells more effectively than parental cells. Hence, these molecules are termed MDR selective (MDRselec) agents. Compounds with a ratio of ≤1 inhibit cell growth in both cell lines with almost equivalent potency, indicating that their cytotoxicity is not modulated by the presence of ABCB1 (non-MDRselec). This categorization scheme yielded in total a data set of 86 compounds which is constituted of 37 (43%) MDR-selective agents and 49 (57%) non-MDR selective agents. The data set is provided in the Supplementary Material accompanying this contribution (Supp SS1). Structures are available as SMILES codes.

Calculation of chemical predictors

The Mold^2 software was used to calculate chemical descriptors that served as numeric representations of the chemical information in the data set. The Mold^2 software enables the generation of an information-rich and comprehensive descriptor set, which in total comprises 777 2D chemical predictors [45]. Mold^2 is a free-of-charge software package, which has been shown to convey a similar amount of information than commercial descriptor calculation

software. Furthermore, Mold^2 descriptors gave better results in benchmarking studies than other predictors. The 777 descriptors available in this package can be grouped into 20 different categories, such as simple atom or functional group and property counts, physicochemical properties, topological and distance indices, autocorrelation vectors, eigenvalue-based descriptors and connectivity indices. The Mold^2 executable was obtained from the FDA webpage [46].

Train/Test Set splitting

In this study we employ a a-priori defined test set for external model validation. The data set presented by Türk et al. consists in total of 86 compounds. In order to generate a representative test set we applied the maximum dissimilarity approach proposed by Willett [47]. The underlying principle of this approach is as follows: Given a data set A with m samples and a larger set B constituted of n samples, it is the idea to generate a sub-sample of B that is diverse to A. To fulfill this task, for each sample in B the m dissimilarities between each point in A are calculated. The most dissimilar point in B is added to A and the process continues until a predefined stopping criterion is reached. In our case we first determined the centroids of each of the two classes in the 777-dimensional input space using pam clustering. These two centroids served as starting points for the generation of the training set. After that the most dissimilar compounds to these two centroids were iteratively added from the data set of 86 compounds until a training set of 80% (stopping criterion) of the molecules was constructed. This procedure yielded a training set of 68 (80%) compounds. The remaining 20% of compounds represented the test set. The properties of these sets are shown in Table 1. A visual representation of the distribution of the training and test set is given in 2-dimensional PC space in Fig. 2.

Table 1: Data set characteristics

	# molecules		physicochemical properties			
	# MDRselec	**# non-MDRselec**	**nHBA**	**nHBD**	**SLogP**	**MW**
training set	30	38	3.5(1.9) [3]	1.8(0.9) [2]	2.3(1.7) [2.3]	316.2(87.6) [312]
test set	7	11	3.1(0.5) [3]	1.1(0.6) [1]	2.6(1.7) [2.6]	352.9(97.1) [307]
total	37	49	3.2(1.9) [3]	1.6(0.9) [2]	2.4[1.7] [2]	328.8(100.2) [308]

Variable Selection by unsupervised forward Selection (UFS).

We applied the unsupervised forward selection (UFS) algorithm proposed by Whitley [48], which represents a versatile tool to reduce oversquare matrices to a size for which simple, robust and easily interpretable inputs for model generation are produced. UFS selects predictors by starting with those two variables that are least well correlated and subsequently selects additional variables on the basis of their multiple correlation with those already chosen, thus generating a feature subset that is close to orthogonality. In order to fulfill this task UFS is comprised of eight different steps that are explained briefly: *(I)* At first a n x p descriptor matrix $X = (x_{ij})$, where x_{ij} is the value of the *j*-th descriptor for the *i*-th compound in the descriptor matrix is mean-centered along its columns. *(II)* Secondly all descriptors that show a smaller standard deviation than the predefined ε-criterion are rejected. *(III)* Next the remaining variables are normalized to unit length. *(IV)* In the fourth step the correlation matrix $(r_{ij}) = X^T X$ is calculated and those two descriptors that show the smallest squared correlation coefficient (r_{ij}^2) are selected and variables whose r_{ij}^2 with either of the two exceeds the predefined R_{max}^2-criterion are rejected. *(V)* Next, an orthonormal basis for the subspace spanned by the first two columns is calculated using the Gram-Schmidt procedure to compute the QR decomposition of this matrix. *(VI)* R_j^2 for all the remaining descriptors with the selected descriptors are calculated and *(VII)* those descriptors that show higher R_j^2 values than predefined R_{max}^2 are rejected. *(VIII)* At last, the matrix with the selected is again orthogonalized and the algorithm returns to step *VI*. Steps *VI* to *VIII* are repeated until all descriptors are either selected or rejected. In our implementation the two hyperparameters ε, which controls the minimum information content of each descriptor and the R_{max}^2 value, which is responsible for the multicollinearity of the selected variables were chosen to take the values ε = 0.05 and R_{max}^2 = 0.75. The source code for the UFS algorithm was implemented into a function in R language and is available from the Supplementary Information accompanying this paper.

Modeling Algorithms.

In this study we employ three different machine learning algorithms to construct classification models of the data used. All three modeling tools belong to the class of supervised learning methods and comprise Leo Breiman´s Random Forests (RF) algorithm [49], a Gradient Boosting Machine (GBM) [50, 51], and a support vector machine (SVM) [52, 53]. The first two algorithms represent so-called ensemble machine learning methods. Ensemble algorithms are basically defined as a set of weaker (so-called "base learners") models that are usually built on a subsample of the whole training space, in an effort to generate a single ("strong learner") model [54]. Ideally, the base learners of such an ensemble complement one another and each one is capable to correctly classify a reasonable amount of the training data [55]. SVM differs from the first two algorithms since it is not an ensemble algorithm, hence its calculated decision boundary is based on the whole set of training instances. All three algorithms have in common, that they have been proved effective for classification of non-linearly separable data [56]. The theory of these ML algorithms is extensively described in the literature [55, 57-60]. Here, only the basic concepts of these methods with a focus on training hyperparameters and certain features that were necessary to produce the results of this study are outlined briefly.

Random Forests (RF). RF belongs to the class of ensemble algorithms and was introduced by Breiman in 2001 [49]. As its names implies, RF combines a collection of

decision trees ("Forests") to produce a more robust and accurate classifier. Decision trees establish classification rules by recursive binary partitioning into nodes that are increasingly homogeneous with respect to the dependent variable [57]. The resulting tree-like structure of such a machine learning is in most cases intuitive to understand and can easily be interpreted. However, a particular disadvantage of such classification trees is that they are highly sensitive to small perturbations in the data and therefore often perform poorly on external test sets or real-life applications. Hence, the generation of many, highly diverse decision trees that are finally gathered to construct one model, is the realization of the RF concept. RF constitutes a collection of *B* trees $\{T_1(X), \dots, T_B(X)\}$, where $X = \{x_1, \dots, x_p\}$ represents a *p*-dimensional vector of predictors associated with a training instance. The ensemble returns *B* outputs and these outputs are aggregated to result in one final classification. For classification problems the final output is returned in form of a proportion of votes for a given compound. For example, a final output for an active molecule of 0.80 means that this compound was classified as an active molecule in 80% of the decision trees of the ensemble. In this study we use a proportion of votes > 0.5 to assign a predicted class label. The training procedure of RF is constituted mainly of two steps: *(I)* First, a bootstrap sample (random sampling with replacement) is drawn from the training instances and subsequently *(II)* a classification tree is grown without tree pruning for this sample using the CART implementation for tree induction [61]. These two steps are repeated until *B* trees are constructed. A particular feature of RF is that it only uses a predefined number of descriptors (m_{try}) for tree induction, rather than using all features. m_{try} essentially represents the only hyperparameter that requires tuning. The default value of m_{try} is $\sqrt{n\ features}$. In this study we select m_{try} to be {2,16,30}, respectively. For the sake of completeness it is to mention that RF provides some type of internal validation during its training routine. Since each tree is induced by only a subsample of the training data, some of the instances are "left out". These instances are referred to as the Out-of-Bag (OOB) compounds. These compounds can readily be predicted by this particular tree, because they have not been used in the tree induction process. This simply means that RF on principle requires no explicit type of cross-validation or a similar method for assessing training performance. However, in order to provide comparability between the other two ML algorithms employed in this study we do not make explicit use of the OOB predictions during training. However, the OOB concept is not only necessary for parameter tuning, but is also responsible for the calculation of RF`s measure of descriptor importance that can be used for subsequent model interpretation. RF calculates the importance of descriptors in the following way: For each tree OOB predictions are made. Simultaneously, each variable in the OOB set is randomly permuted, one at a time, and each of these modified OOB sets is additionally classified by the tree. At the end of training, the margins for each molecule are calculated on basis of the OOB results as well as on the permuted OOB results. For a two class problem the margin is calculated as the difference between the proportion of votes for the correct class and the proportion of votes for the incorrect class. Finally, the descriptor importance of the *j*-th descriptor is: $M - M_j$, where M is the average margin based on the OOB predictions and M_j is the average margin based on the permuted OOB predictions. Another appealing feature of RF is that it additionally provides an intrinsic proximity measure that can be used for visualization of data in descriptor space, clustering of molecules or for molecule

neighborhood assessment. This proximity measure is calculated as follows: A set of molecules whose pairwise proximity is of interest are predicted by RF. The proportion of trees in the ensemble where a pair of molecules landed in the same classification node constitutes the proximity between these two molecules. Conclusively, the proximity between a compound and itself is always 1. Obviously, this type of proximity measure has one striking advantage. Contrary, to other similarity metrics it is computed in a supervised manner, since the activity of interest dictates tree structures in the Forests. It needs to be mentioned that RF can also be run in an unsupervised mode. We use this proximity measure in conjunction with multidimensional scaling to visualize our data set and we compare it to a commonly used distance metric, namely the Euclidean distance (see Figure 7). Additionally, the RF implementation in R also allows a detailed analysis of the effects of certain descriptors on the classification performance by means of partial dependence plots. Partial dependents plots are graphical visualizations, which display the distribution of a descriptor on the x-axis and show the logit of probability (i.e. the log of the fraction of votes for a class) of a certain class calculated by the RF model on the y-axis.

Gradient Boosting Machine (GBM): Boosting was first mentioned by Freund and Shapire with the publication of the AdaBoost algorithm [55, 57, 62]. The idea was to combine simple tree-based models to an ensemble in such a way that the performance of any ensemble member exceeds (is "boosted") the performance of its predecessor. In comparison to the RF algorithm boosting is an iterative process that weights the contribution of each base learner by its performance rather than assigning equal weights to all models. Another difference to RF is that boosting in general is not algorithmically constraint, whereas a RF ensemble can only be constructed out of decision trees [63]. While the first implementations of boosting could not take full advantage of the base learners, further advances that utilize the concept of iterative, stochastic gradient optimization of a loss function $L(y, F(x))$ have led to the currently used GBM methods [51, 64]. Acknowledging the stochastic nature of these machines, they are often referred to as stochastic GBMs (SGBM or SBM). Throughout this paper we treat these terms equivalent. The GBM training procedure builds a sequence of M decision trees in a stage-wise manner out of a given training set of N compounds. The loss function $L(y, F(x))$, which is minimized at each step m represents the degree or wrongness of a base learner at m. In classification $L(y, F(x))$ is described as the binomial log-likelihood function $L=\log(1 + \exp(-2YF(x)))$. At each iteration step m, a decision tree is fitted to a randomly selected subsample s of the gradients of the loss function. This random sampling perturbates the data in order to maximize the diversity and simultaneously reduces the bias of each base learner. The bias of each base learner can be further reduced by controlling the number of splits that is used in the generation of each decision tree m. Trees with many splits comprise a lower bias than those trees that are constructed using relatively few splits. This hyperparameter is termed interaction.depth [65]. We evaluate this hyperparameter by using different values {1,2,3} during training. Another hyperparameter of GBM that requires tuning is the number of trees (ntrees) that is used to constitute the final ensemble. In this study we use values {50,100,150} for ntrees. The shrinkage parameter v, which controls the learning rate, is held constant during training using a value of 0.1 as recommended in the literature [65]. Additionally, the GBM method also provides a measure of variable importance, which

can be used in the course of model interpretation. Briefly, variable importance is computed as the relative influence of each variable in reducing the loss function at each iteration. The GBM implementation in R utilizes the AdaBoost algorithm [65], which is notorious for its exponential loss function described above, but uses Friedman´s gradient descent algorithm for adjusting the learning rate, rather than the original one proposed in the AdaBoost algorithm.

Support Vector Machine (SVM): The main motivation of SVM learning is based on the maximum-margin hyperplane concept introduced by Vapnik [53]. According to this, SVM generates a hyperplane which separates data points of two classes which are represented by a vector x_i of compounds with a maximum margin. The instances that define that margin, i.e. are located closest to the hyperplane, are called the support vectors [57]. This is of particular importance, because once a hyperplane is found only the support vectors are used to define the solution as linear combination of only these instances, while remaining data points are ignored [56]. An important consequence of that is that the model complexity is unaffected by the number of descriptors in the data set. SVM training in its basic form is done by finding a vector w and a constant b that minimizes $||w||^2$ and satisfies $w * x_i + b \geq +1$ for the positive (active) class and $w * x_i + b \leq -1$ for the negative (inactive) class. Here w is a vector perpendicular to the hyperplane, $|b|/||w||$ is the orthogonal distance from the hyperplane to the origin and $||w||^2$ is the Euclidean norm of w. As soon as w and b are known, a given compound x_i can be classified by the following decision function: $f_{w,b}(x) = sign[(w * x) + b]$. For many data sets, SVM may be unable to find an optimal hyperplane, because certain training instances are always misclassified. In order to deal with that situation a *soft margin* is used that accepts some misclassifications [66]. This *soft margin* is controlled via the hyperparameter *Cost (C)*. *C* controls the trade-off between allowing misclassifications in training and forcing rigid margins. Noteworthy, with increasing the value of *C*, forces the generation of a rigid margin and a putative more accurate model, that may suffer from overfitting. In order to evaluate the optimal value of C during training we use for C={0.1,1,10}. Furthermore, this simple scheme of SVM learning is only appropriate if data are linearly separable. However, in many application areas, data cannot be separated linearly. In such cases, SVM maps the input feature space into a much higher dimensional feature space, with the aim of making the class separation easier in that space. It is important to note that a linear separation in transformed feature space corresponds to a non-linear separation in the original input space [67, 68]. This mapping is done by applying a kernel function $K(x,y) = \Phi(x_i) * \Phi(x_j)$ [69]. According to this definition, a kernel function can be generalized as a dot product (or similarity measure) in some feature space. The idea behind the "kernel trick" is that Φ does not need to be calculated, since kernels allow inner products to be calculated directly from feature space [68]. In this study we apply a commonly used polynomial kernel $K(x,y) = (\varphi * x * y + 1)^p$ [70]. Two important parameter of this kernel are tuned during training. One is the degree of the polynomial function *p*. We select $p = \{1,2,3\}$ during training. And the other one is the scaling factor φ. For φ we select values of {0.001, 0.01, 0.1}. Usually, SVM reports a binary class label for classification of new instances. We make use of an improved implementation that enables us to retrieve class probabilities for

each compound (we apply a probability threshold of ≥ 0.5 for assigning compounds to the positive class).

Table 2 gives an overview on the properties of the discussed ML algorithms and also summarizes the hyperparameters for each algorithm.

Table 2: Characteristics of the ML algorithms employed in this study

	Random Forests (RF)	Gradient Boosting Machine (GBM)	Support Vector Machine (SVM)
ensemble method	yes	yes	no
base learner	tree	tree	none
unsupervised mode	possible	no	no
linear model	no	no	yes
tools for interpretation	yes	yes	no
ease of parameter handling	***	**	*
tuning parameters	m_{try}	interaction.depth, ntree	*C*, scale, *p*

Model training, performance measures and model selection.

During the generation of a ML model, the key points to verify are optimal setting of hyperparameters and model robustness, which in case of a classification model can be verified by various parameters using internal validation techniques. Internal validation is usually carried out by iterative data set splitting. 10-fold crossvalidation is among the best known and widely used methods. Here, we employ a grid search strategy for training of the models using three independent repetitions of 10-fold crossvalidation. To estimate the training performance indices derived from the confusion matrix can be used. The confusion matrix arranges the number of true positivies (TP), the number of true negatives (TN), the number of false positives (FP) and the number of false negatives (FN) in a symmetric matrix. Accuracy (Acc=(TP+TN)/(TP+TN+FP+FN), sensitivity (Sens=TP/(TP+FN)) and specificity (Spec=TN/((TN+FP)) are widely used performance measures computed from the confusion matrix [71]. However, none of these can be reliably used on its own, since they harbor several disadvantages that could bias model selection. For instance, accuracy depends on the class distribution of the training data and the other two parameters are only class specific [38]. A better alternative is to use the Receiver Operating Characteristic (ROC) chart, which plots the sensitivity (also known as true positive rate) against the false positive rate (1-specificity) [72, 73]. ROC curves visually represent the information of a confusion matrix in a much more intuitive and robust fashion [74]. Furthermore, the Area under the ROC curve (AUC), can be readily calculated for any classification model that reports a probabilistic classification result

[75-77]. The AUC reports the probability of molecules of interest (e.g. actives) being ranked earlier than decoy compounds (e.g.: inactives) and can take values between 1.0 (perfect model) and 0.5 (random model). In this study we use the mean AUC as decision criterion for model selection in the three repetitions of 10-fold crossvalidation.

Assessment of variable importances.

For RF and GBM models we used the model-specific measures of variable importance as described above. Additionally, we also made used of model-independent metrices, which are available in R. For the model-independent metric, ROC curve analysis was conducted on each descriptor. A series of cutoffs was applied to the descriptor data. Consecutively, sensitivity and specificity were calculated for each cutoff and the area under the ROC curve is computed using the trapezoidal rule. This area was used as a measure of variable importance.

Software.

Chemical structures were downloaded from the pubchem database. Molecular features were calculated using the free-of-charge software package Mold2 provided from the FDA. Calculation of physicochemical parameters was carried out using the Rchemometrics package. The calculation of dissimilarities in the course of the train/test splitting was done using the proxy package available in R. The maxDissim()-function was used to generate the training set. All machine learning models were generated using the R software environment. The caret package was used to construct Random Forests, GBM and SVM models. Principal component analysis was conducted using the prcomp()-function in R and the covariance matrix was used to calculate eigenvectors and their respective eigenvalues. For multidimensional scaling the isoMDS()-function available in the MASS package in R was used.

Results and Discussion

Preliminary data characterization and data preprocessing.

At first we conducted a preliminary examination of the data set using a linear discriminant analysis and a decision tree as classification tools with the aim to receive easy interpretable models that characterize MDR-selective compounds. Unfortunately, the retrieved results were disappointing, since the models gave only a poor training performance (results are provided in Supplementary Information ST1). This prompted us to use more sophisticated methods to unravel the obvious complex and supposedly non-linear SAR of this data. The modeling workflow that we conducted in this study is visualized in Fig. 1.

We started with computing an information rich and high dimensional descriptor set using the 777 2D descriptors implemented in the Mold2 software package. Subsequently, we split our data set into a training and a corresponding test set using a dissimilarity-based algorithm. This procedure resulted in a training set constituted of 68 compounds (30 MDR selective and 38 non-MDR-selective) and a test test of 18 compounds (7 MDR-selective and 11 non-MDR-selective). The characteristics of the training and the external test set alongside with classical physicochemical properties are given in Table

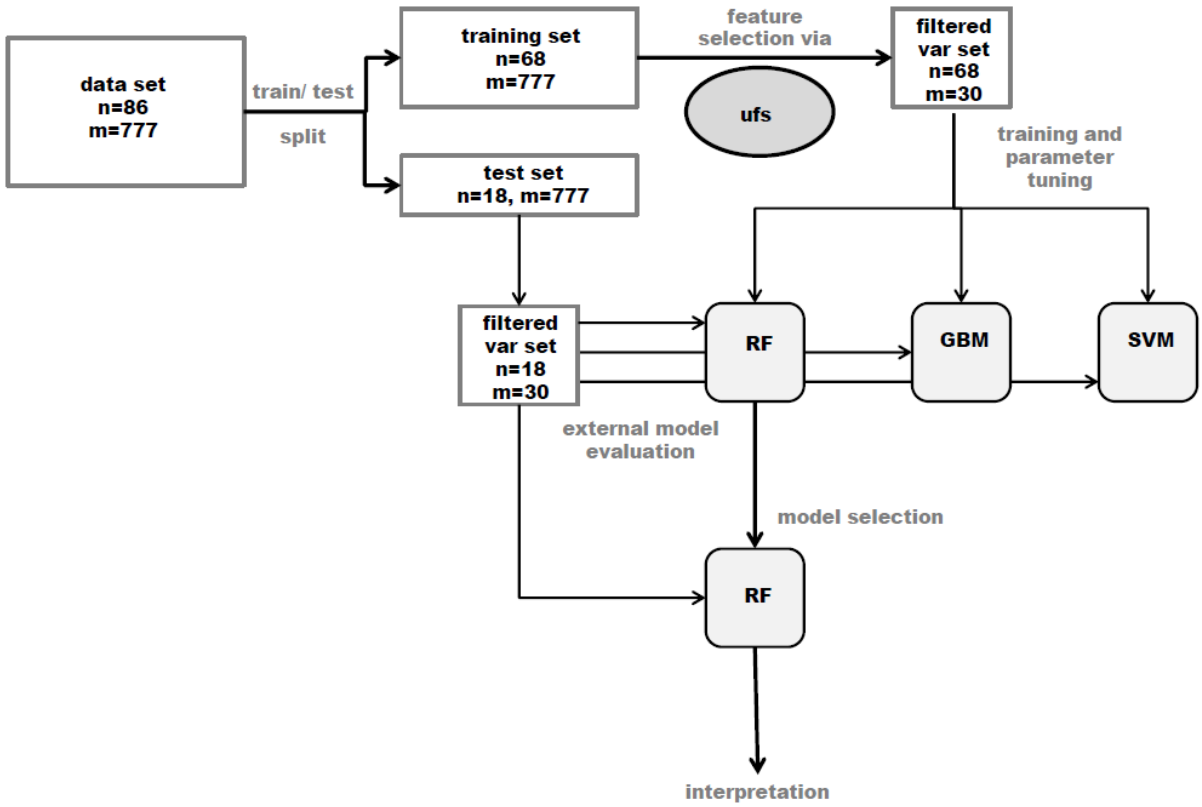


Figure 1: Modelling Flowchart employed in this study

1. It can be seen that these basic properties are evenly distributed between the two sets. Furthermore, the chemical space spanned by the training and test set by means of a principal component analysis (PCA) is represented in Figure 2. Symbols denote the different data sets. Figure 2 resembles the first two axes, which explain a cumulative variance of 64.8% of the data. It can be seen that training and test set populate the same space and are highly overlapping. This further supports, that descriptor information is evenly distributed among the sets and that our procedure yielded an unbiased and representative training and test set. According with our flowchart, we went on with the application of Whitley`s UFS algorithm.

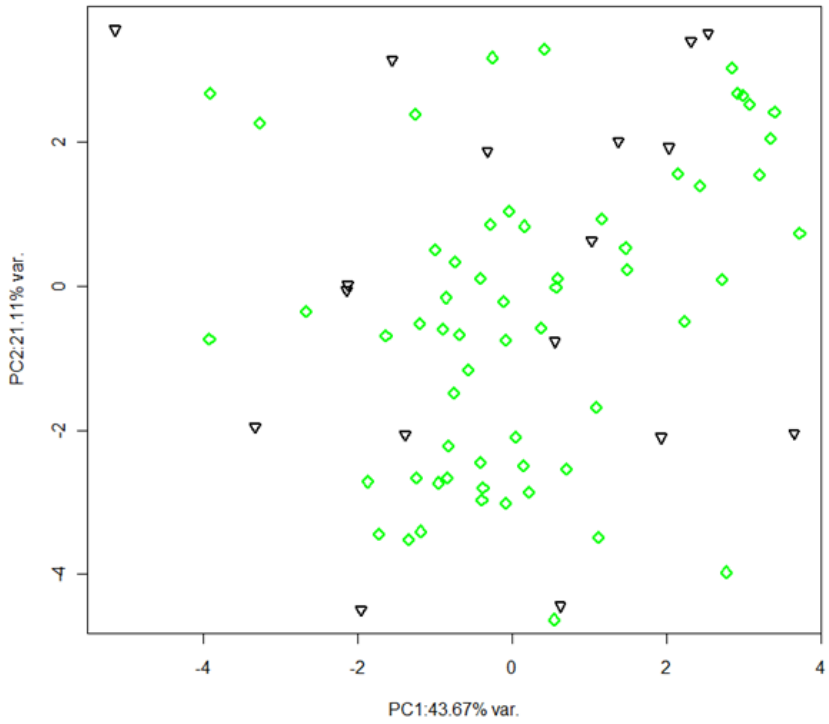


Figure 2: Principle Component Analysis showing the distribution of training and test set compounds in Mold2 decriptor space.

The primary motivation for the use of UFS was to achieve a descriptor matrix that resembles mostly relevant and non-redundant features that additionally comprise only a minimum amount of multicollinearity. As outlined in the methodology section UFS conducts feature selection in a two-step manner. The rejection of irrelevant descriptors that contain only a small degree of variation resulted in a data set of 523 descriptors and the subsequent selection of descriptors that show only a small degree of intercorrelation yielded a final descriptor set of only 30 descriptors. The training matrix being built of the 68 molecules and the 30 selected features was subjected to the three ML algorithms (RF, GBM, SVM) employed in this study.

Parameter tuning and model selection.

In order to develop highly accurate classification models, careful tuning of algorithmic specific hyperparameters in the course of the internal validation procedure is essential. For internal validation we conducted three repetitions of 10-fold crossvalidation (CV) on our training set. The results were averaged and the model that showed the area under the ROC curve value closest to 1.0 was selected for further external validation. The results of model training are graphically depicted in Fig. 3A-C.

RF is the algorithm with only one hyperparameter to tune, namely mtry, which sets the number of variables used at each split in the course of tree building. This parameter was gradually decreased from 30 to 16 to 2 resulting in a total number of three models. From Fig. 3A it can be seen that the RF model with the best training performance is obtained with mtry being set to 16. This model gave an average ROC of 0.88 and a SD of 0.05 in the three

repetitions of CV. The GBM algorithm offers two parameters requiring tuning in the course of training. These are the number of trees (n.trees) used for the whole ensemble and the number of splits (interaction.depth) that are carried out in each tree. In total 9 GBM models were generated during the grid-based parameter optimization procedure. Fig. 3B reports the results of the grid-based training procedure carried out to estimate the best values for the two parameters. It is shown that the best performing model with a ROC of 0.862 (SD±0.06) is obtained using 150 trees to construct the ensemble with each tree having an interaction.depth of 3. With respect to the number of tuning parameters SVM is definitely the most complex algorithm used here. For SVM we evaluated three different parameters: two kernel-specific parameters; and one kernel-independent parameter. In total 27 SVM models with different parameter settings were generated. The evaluation of these parameters showed that the SVM model using a polynomial kernel constructed with a polynomial degree of 2, a scaling factor of 0.01 and a C of 1 returned the best training results (Fig. 3C).

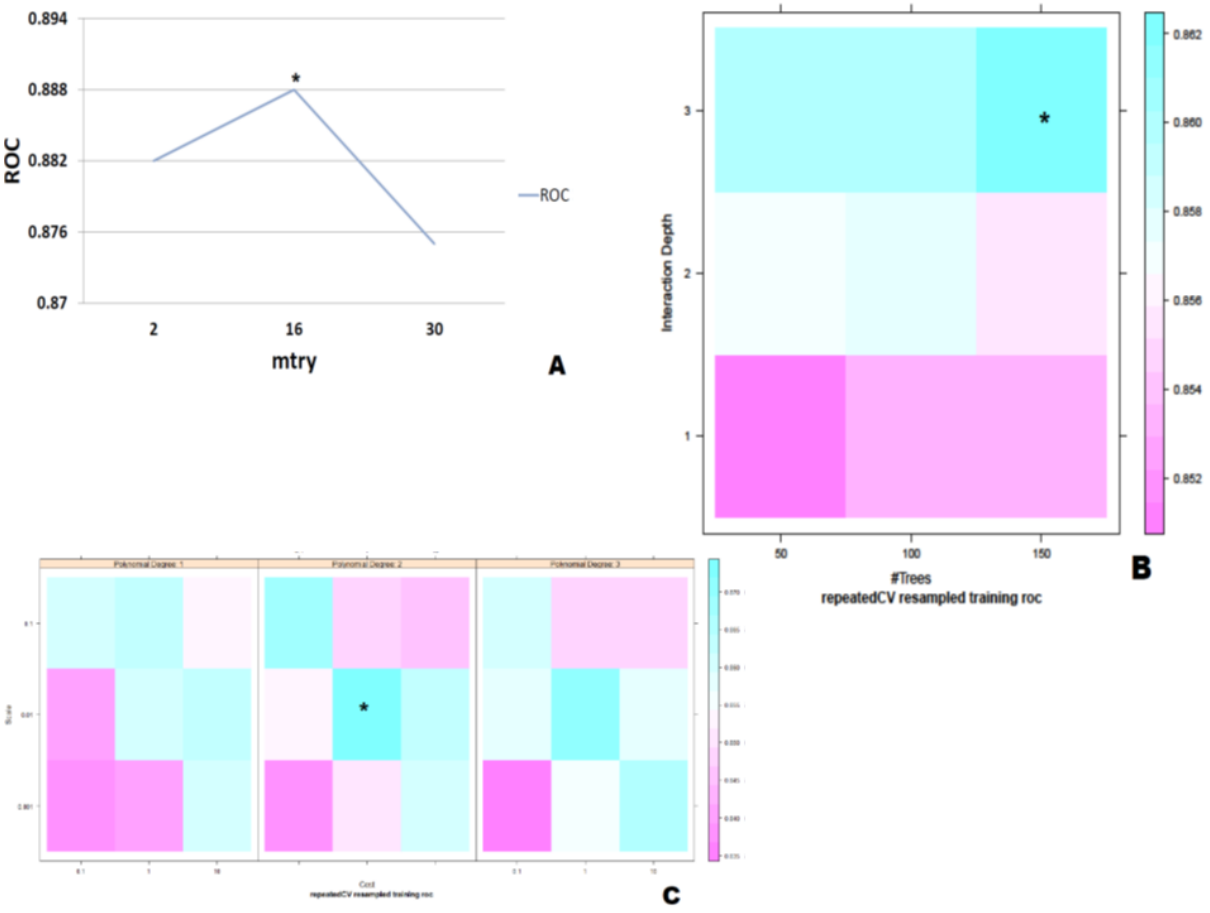


Figure 3: A-C): Parameter tuning of the three modeling algorithms in three independent runs of 10-fold cross-validation

This model showed an average ROC of 0.872 (±0.04). The full list of all the different training optimization runs carried out for the three algorithms including performance measures can be found in Supplementary Tables ST2-ST4. Summarizing the results on our training set it can be said that all three algorithm revealed good performance with ROC values always higher than 0.85. However, comparing the best models for these three algorithms in more detail suggested that RF might be most suited for external prediction since it gave the most robust

training performance. SVM gave the second best performance with respect to ROC performance and GBM performed worse than the others. The best models including sensitivity and specificity measures computed from the confusion matrices are reported in Table 3.

Table 3: Mean performance measures of the best models for each modeling algorithm in the three repeated 10-fold cross-validation runs.

Model	Sens	Spec	ROC	Sens SD	Spec SD	ROC SD
RF	0.919	0.743	0.888	0.0534	0.0987	0.048
GBM	0.874	0.733	0.862	0.0784	0.104	0.0601
SVM	0.893	0.727	0.872	0.0638	0.0988	0.0491

Assessment of model performance in terms of external predictivity.

Additionally we utilize the externally derived test set consisting of 18 molecules to evaluate our three best performing models. The resulting statistical parameters are displayed in Table 4. Again, the model derived using RF as classification algorithm gave the best performance (ROC= 0.79). Overall, the RF model classifies 83% of the compounds correctly. It was able to correctly assign 10 of the 11 non-MDR-selective test set compounds and also correctly classified 5 of the 7 MDR selective compounds, which resulted in a specificity of 0.91 and a sensitivity of 0.71 respectively. The other two models showed a weaker classification performance on this external test set.

Table 4: Confusion Matrix and performance measures of the best performing classification models in the prediction of the external test set.

Model	TP	TN	FP	FN	ROC	Acc	Sens	Spec
RF	5	10	1	2	0.79	0.83	0.71	0.91
GBM	4	9	2	3	0.74	0.72	0.57	0.82
SVM	4	7	4	3	0.66	0.61	0.57	0.64

It is remarkable that despite the excellent specificity values retrieved for the GBM and the SVM model during training, they are only able to correctly assign 4 out of the 7 MDR-selective molecules when challenged with the external test set (Sens=0.57). Nevertheless, the GBM model performs much better than the SVM model with respect to the correct assignment of the inactive class. 9 out of 11 inactives are correctly predicted by the GBM model (Spec=0.82), whereas the SVM model is only able to classify 7 of these molecules accordingly (Spec=0.64). Figure 4 shows the ROC plot for the three models when predicting the external test set. The black diagonal ranging from the coordinates [0,0] to [1,1] represents the performance of a random model, which is not able to distinguish signal from noise. Consequently, for any possible threshold the same percentage for sensitivity (signal) and 1-specificity (noise) is achieved. The AUC of the diagonal is 0.5 and any classifier above that limit can be considered as being better than random. The ROC curve of an ideal classifier harbours an AUC of 1 and coincides at the coordinates [0,1] at the top left edge of the plot. From Fig. 4 it can be seen that the three models behave quite differently. From the perspective

of a virtual screening situation the left part of the plot is most important, because it denotes how much signal (here MDR-selective agents) can be identified by the model while still discarding most of the noise. In the case of the GBM model almost 40% of MDR-selective compounds can be retrieved without classifying any false positives. At this part of the plot GBM outperforms the other two models. However, at approximately a true positive recovery rate of 50% GBM looses its superiority and RF catches up to finally report the best classification result. Again, the weak performance of SVM is clearly evident from the ROC plot. Compared to their training performance all three models show weaker performances, but RF still outperforms the other algorithms as suggested by the training results.

However, the GBM and the SVM gave much weaker performances in external evaluation, which is indicative of "overfitting", a phenomenon that describes excellent training performance, but poor test set performance. For the SVM model the weak test set performance can be explained by its algorithmic nature. SVM classification is based on a linear decision boundary that is usually constructed to optimally reflect the training data, which by definition is a potential source for overfitting. Although, we accounted for that by trying to carefully select the kernel-independent soft margin hyperparameter C it appeared to be inefficient to correctly predict the test set.

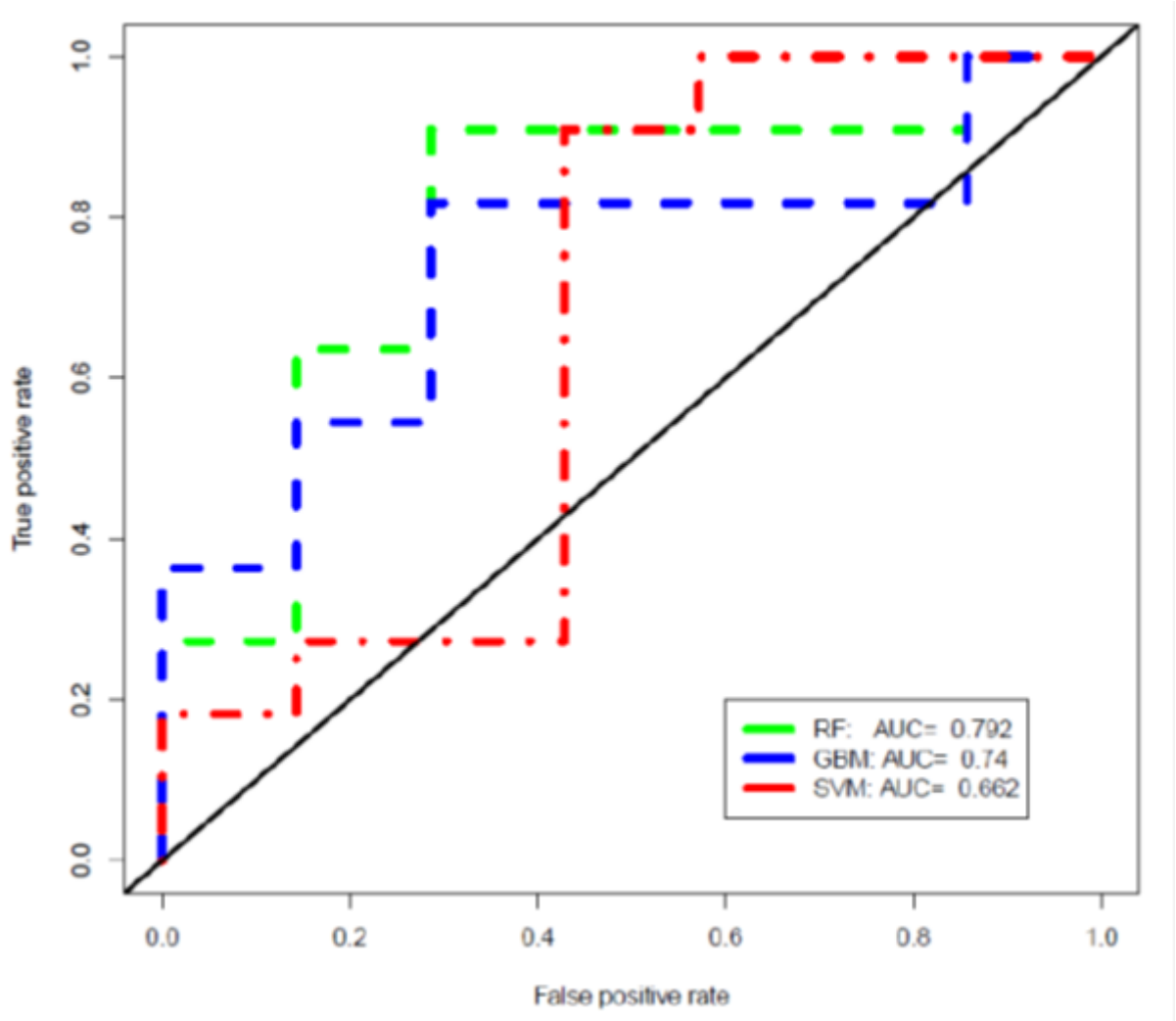


Figure 4: ROC curves of the prediction performance of the external test set.

Furthermore, in comparison to RF and GBM, which clearly outperform SVM on the external test set (at least with respect to specificity), SVM is a single learner that constructs its classification boundary once on all the training data, whereas the other algorithms represent ensemble methods, that carefully mine different constellations of the training data and base their final classification result on different models generated on various subsamples of the training data. An explanation why GBM does not reflect a similar training performance when challenged with the external data set is less intuitive. Given the fact that GBM is also an ensemble method like RF it should be resistant towards overfitting. However, when recalling the differences between GBM and RF it becomes apparent that GBM utilizes a learning rate during construction of base learners. We have to admit that we held this learning rate constant in course of training and focused on the more obvious hyperparameters, ensemble size and tree size. Additional tuning of this parameter might possibly modify GBM performance. These results clearly highlight that exhaustive internal validation and parameter selection do not guarantee generalisability of ML models and further undermine the necessity of using external data sets for model evaluation. Summarizing, the classification results presented here, it can be generalized that our results with respect to performance of the different algorithms are consisted with previous findings [57, 78, 79]. In general, ensemble methods

outperform single learners and RF is the most suitable algorithm for modeling complex relationships. A further important aspect of QSAR model validation is to estimate if the best model has arisen by the play of chance alone. An efficient way to estimate putative chance correlations is to utilize Y-randomization. In order to perform Y-randomization the Y-variable, i.e. the class label of the training set was randomly permuted. Based on this permuted set a RF model was regenerated using the same tuned parameter (mtry=16) and the resulting model was used to predict the external test set with the original Y-variable. This procedure was repeated 100 times and the resulting sensitivity and specificity measures are graphically visualized in Fig. 5A.

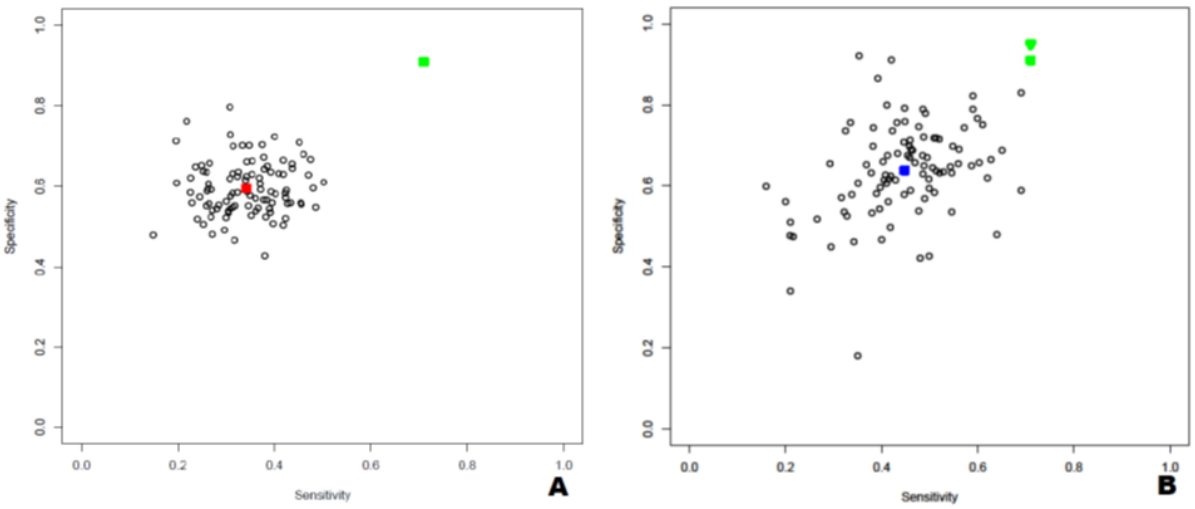


Figure 5: A-B: Scatterplot showing sensitivity and specificity.

Y-randomization results clearly rule out that the RF model has resulted from chance alone. The 100 generated models are in line with what is expected on the basis of mere random classification. On average these 100 models report a sensitivity of 0.37 and specificity of 0.59, whereas the best RF model of the actual data has a sensitivity of 0.71 and a specificity of 0.91. Another important aspect of model validation is to assess the classification capability of the used descriptor. In our case this is of special interest since we selected our descriptors on a solely unsupervised basis. To assess this issue, we proceeded in a similar fashion as for the Y-randomization experiment, but instead of randomly permuting the response variable we randomly picked the same amount of descriptors (n=30) and generated a RF model that was subsequently used for classification of the test set. Again this procedure was repeated 100 times. The results are depicted in Fig. 5B. This plot also depicts the sensitivity and the specificity measures on the axes. It can be seen that the models constituted on randomly picked variables show a much larger distribution with respect to these two performance measures when compared to the performance in Y-randomization. Additionally, there are also some models that either achieve a similar sensitivity or a similar specificity than our model based on the 30 UFS selected descriptors (green square in Fig. 5B). Nevertheless, none of these models showed a similar performance regarding both performance measures. This clearly highlights that feature selection on basis of the UFS algorithm indeed results in a feature set optimal for classification of MDR-selective compounds. Furthermore, we also provide information on the RF model constructed out of all 777 descriptors in Fig. 5B (green

triangle). It is shown that specificity can be increased using all this features, but nevertheless sensitivity is unaffected. Given the fact that a model based on 777 descriptors is much more complex and more difficult to interpret, it can be summarized that a UFS-derived feature set is suitable with respect to its size, but also with respect to its predictive power. In an attempt to appraise the class separating property of the RF model based on the 30 UFS descriptors we utilize RF´s built-in proximity measure to compare the distribution of MDR-selective and non-MDR selective molecules in the feature space. We utilized the RF dissimilarity as input for isotonic MDS to arrive at a low dimensional data representation. Fig. 6A and Fig. 6B compare the class specific data configurations in "classical" Euclidean Space (Fig. 6A) to the configuration retrieved by using RF derived proximities (Fig. 6B).

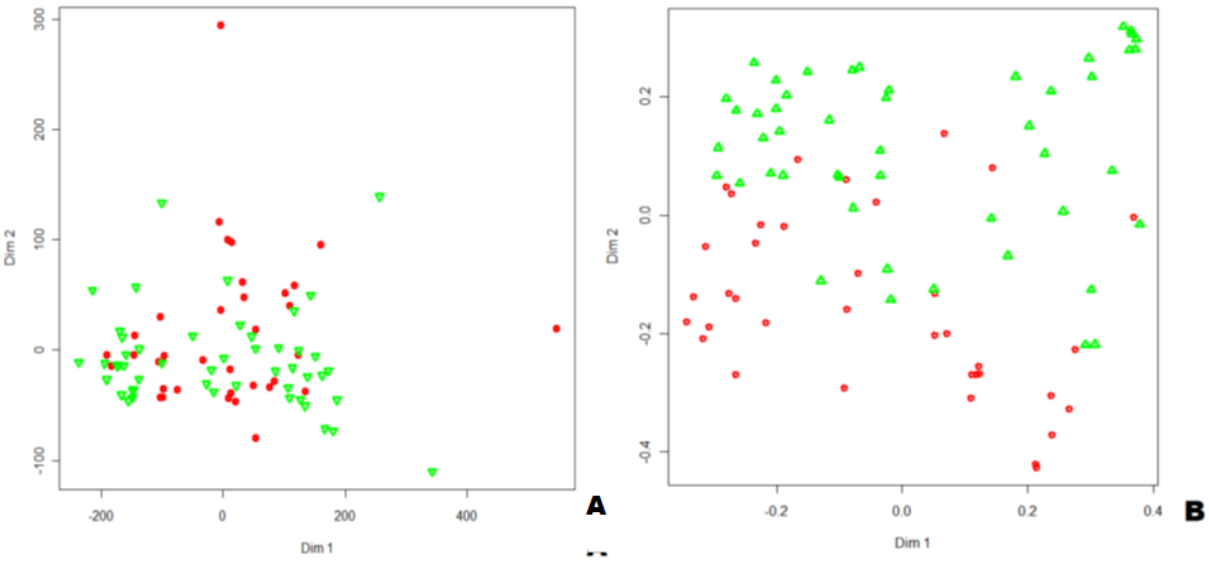


Figure 6: (A-B): Visual comparison of the class separating power of the 30 UFS-selected descriptors using isotonic multidimensional scaling (isoMDS).

The 30 UFS selected descriptors have been used as input for calculating the different proximity measures in Fig. 6. Visual interpretation clearly highlights the class separating power of RF in comparison to the classical Euclidean distance, which further strengthens the good predictive properties of our best model. Taken together, RF modeling by utilizing UFS-selected descriptors gave the best performance.

Model Interpretation and Assessment of Variable Importance.

One important issue of ML QSAR models is their interpretation. Interpretation mainly aims to provide useful information that might support experimental scientists in decision making when designing molecules with better MDR-selective properties. In general, model interpretation is often done by applying certain measures of variable importance that may either be generated during the course of modeling or can be calculated using model independent measures. In Fig. 7A and 7B we report the ranking of descriptor importances in

terms of a scatter plot matrix that shows the pair-wise correlations of the RF, the GBM and the model-independent descriptor ROC as measures of variable importance.

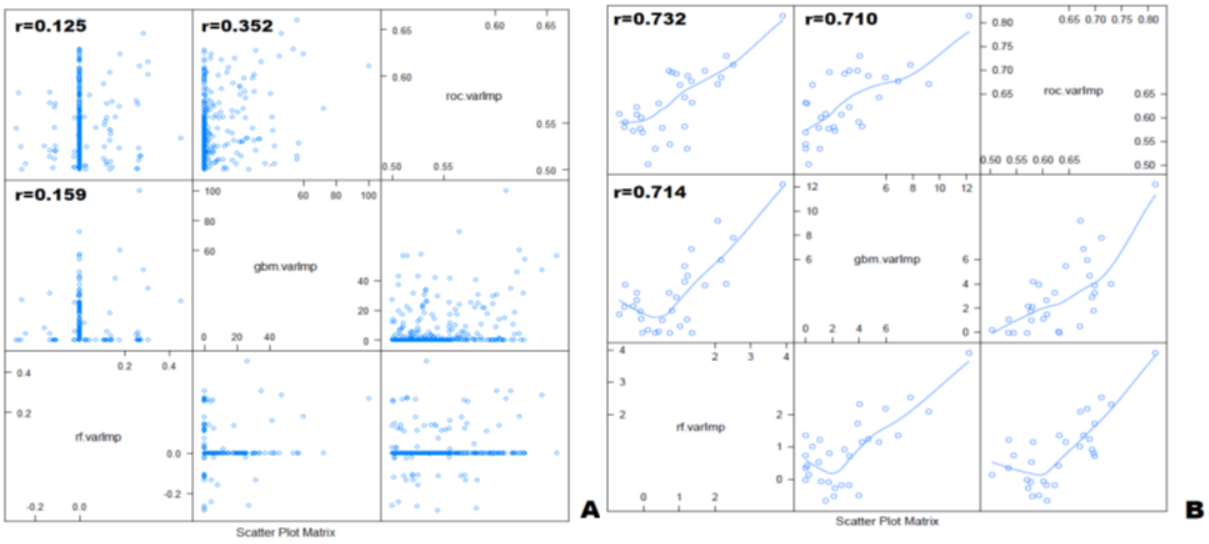


Figure 7: (A-B): Scatter plot matrices showing the correlation of different variable importance measures of the best RF model (rf.varImp), the best GBM model (gbm.varImp), and an additional model independent metric (roc.varImp).

Fig. 7A shows the amount of agreement of the ranking of the three methods when models are generated using all 777 descriptors as inputs, whereas Fig. 7B shows the agreement of the three methods when only using the 30 UFS-selected descriptors. It can be seen that using all descriptors as inputs for model generation results in a lack of correlation (agreement) of the importance measures (in general $r < 0.2$), whereas the UFS-selected descriptors show much better correlations ($r > 0.7$). This shows that the UFS method is capable of identifying relevant descriptors that are nearly equally ranked by different assessment tools of variable importance. Since the RF model with 30 UFS-selected descriptors gave the best performance in predicting the external data set, we restrict our analysis and interpretation to this model. By carefully analyzing the 30 UFS-selected features it becomes apparent, that 11 of these belong to the class of autocorrelation descriptors (D448, D419, D458, D504, D495, D451, D488, D497, D457, D508, D484), 7 belong to the class of functional group counts (D765, D771, D602, D741, D661, D599, D035), 2 denote simple physicochemical properties (molecular weight (D122), molecular regression coefficients surface LogP index (D777)), two describe topological distances between certain atom types (distance between N and P (D367), distance between N and O (D368)) and the remaining describe molecules at a very complex level of sophistication (eigenvalue and path information) and thus are difficult to interpret. Four of the 11 autocorrelation descriptors use the molecular mass as weighting property for descriptor calculation (D448, D419, D451, D484), whereas three use atomic van-der-Waals volumes as weighting property (D458, D488, D457), which indicate that more generally weight and size are important features to characterize MDR-selective agents. Additionally the molecular weight descriptor (D122) is ranked at place 20 according to RF importance, which further undermines the role of molecular weight in the model. Given the fact that the topological distance descriptors (D368, D367) use N for the calculation and that two functional group counts also utilize the presence of N atoms, stresses out that nitrogen atoms are also important contributors to the model. Furthermore, 15 out of the 30 UFS-selected features show p-values below $\alpha=0.05$, which is additionally confirmative of the good classification properties of the selected descriptors. The full list of the 30 UFS-selected descriptors including a short description, their ranking according to the RF model and their p-values calculated using a Mann-Whitney-U test are given in Table 5.

Table 5: Variable importance rankings of the 30 UFS-selected descriptors.

Mold[2] code	Description	RF ranking	ROC	p-value	sig. at α=0.05
D588	highest eigenvalue from Burden matrix weighted by polarizabilities order-1	1	0.598	0.007	*
D448	topological structure autocorrelation length-2 weighted by atomic masses	2	0.646	0.066	NS
D765	number of group R=N-	3	0.621	0.037	*
D301	maximum eigenvalue weighted by heteroatoms and multiple bonds matrix	4	0.660	0.092	NS
D419	topological structure autocorrelation length-4 weighted by atomic masses	5	0.618	0.262	NS
D458	topological structure autocorrelation length-4 weighted by atomic van der Waals volumes	6	0.619	0.022	*
D504	topological structure autocorrelation length-2 weighted by atomic polarizabilities	7	0.628	0.009	*
D771	number of group R=S	8	0.604	0.001	*
D368	sum of topological distance between the vertices N and O	9	0.602	0.008	*
D602	number of ring quaternary C-sp3	10	0.603	0.025	*
D495	topological structure autocorrelation length-1 weighted by atomic Sanderson electronegativities	11	0.558	0.022	*
D572	Highest eigenvalue from Burden matrix weighted by van der Walls order-1	12	0.599	0.018	*
D294	bond information content order-1 index	13	0.565	0.044	*
D195	maximal valence vertex electrotopological positive variation	14	0.503	0.049	*
D741	number of group Ar-CH=X	15	0.582	0.083	NS
D451	topological structure autocorrelation length-5 weighted by atomic masses	16	0.588	0.090	NS
D488	topological structure autocorrelation length-2 weighted by atomic van der Waals volumes	17	0.585	0.189	NS
D497	topological structure autocorrelation length-3 weighted by atomic electronegativities	18	0.609	0.026	*
D661	number of group ammonia groups (aliphatic)	19	0.580	0.023	*
D122	molecular weight	20	0.576	0.078	NS
D353	molecular topological multiple path index of order 05	21	0.601	0.368	NS
D599	number of total quaternary C-sp3	22	0.589	0.041	*
D351	molecular topological multiple path index of order 03	23	0.574	0.241	NS
D457	topological structure autocorrelation length-3 weighted by atomic van der Waals volumes	24	0.563	0.068	NS
D508	topological structure autocorrelation	25	0.557	0.067	NS

	length-6 weighted by atomic polarizabilities				
D591	highest eigenvalue from Burden matrix weighted by polarizabilities order-4	26	0.555	0.011	*
D367	sum of topological distance between the vertices N and P	27	0.527	0.053	NS
D484	topological structure autocorrelation length-6 weighted by atomic masses	28	0.579	0.142	NS
D035	number of Chlorine	29	0.514	0.104	NS
D777	molecular regresson coefficients surface LogP index	30	0.504	0.136	NS

Considering the fact that the numbers of autocorrelation descriptors that are weighted by atomic mass is striking, we wanted to further investigate the distribution of these properties within this data set. Therefore, we used partial dependence plots to graphically characterize relationships between these predictors and the predicted class label (see Figure 8 A-F). Fig. 8 A-D represents the partial dependence function of these descriptors using different bond lengths. Interestingly, for the smaller distances l=2 and l=4 (Fig. 8A, 8B) higher descriptor values show a lower likelihood for MDR-selective compounds (Fig. 8A) or a highly non-linear behavior is observed (Fig. 8B). Contrary to that, for longer distances, which are represented in Fig. 8C and Fig. 8D with l=5 and l=6, an almost linear trend in the opposite direction is observed. In other words, higher values show a higher likelihood for the MDR-selective class, which suggests that MDR-selective agents are larger and show higher atomic masses than non-MDR selective compounds. For the molecular weight descriptor a highly non-linear relationship with respect to the MDR-selective class is observed (Fig. 8E). The partial dependence of lipophilicity, which is shown in Fig. 8F, shows a linear relationship between this property and the active molecules, indicating that less lipophilic compounds are more likely to represent MDR-selective agents. This is of interest, because QSAR studies aiming to describe ABCB1-substrates and inhibitors often suggested, that such molecules are mainly characterized by higher lipophilicity [35, 80, 81]. In summary, it is suggested that the likelihood for designing MDR-selective agents increases with large, heavy molecules that only provide a moderate lipophilicity.

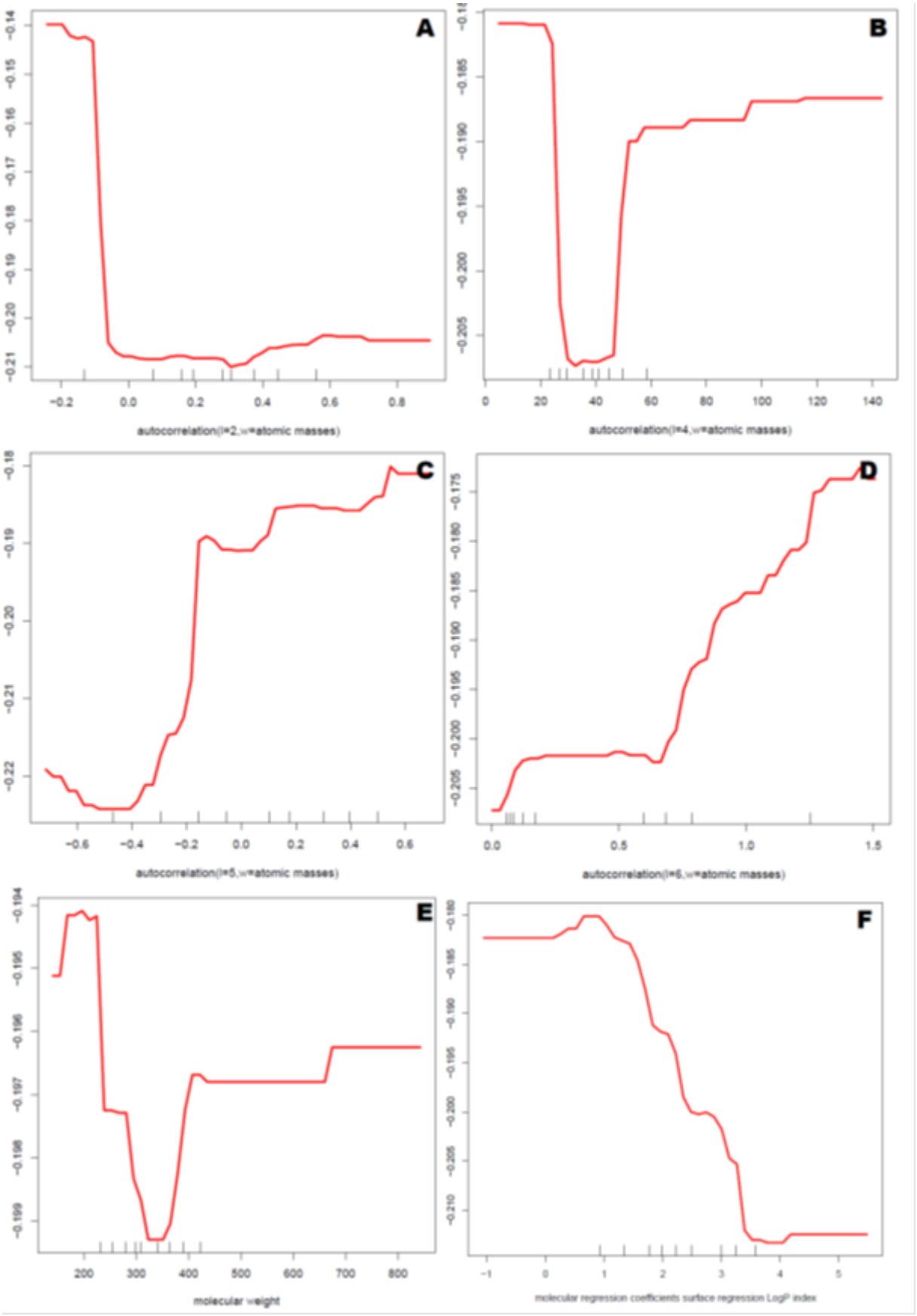


Figure 8: (A-F): Model Interpretation: Partial dependence plot for selected autocorrelation descriptors weighted by atomic masses (A-D); A-D use the same property, but resemble different distances used for computing the

autocorrelation vector. E represents the molecular weight and F shows the lipophilicity. Partial dependence is the dependence of the probability of one class on one descriptor after averaging out the effects of the other descriptors in the model. Therefore, the y-axis of a partial dependence plot represents a logit scale of the MDR-selective class compounds (i.e.; the log of fraction of votes; or more positive values indicate a higher probability of the appearance of the MDR-selective class.)

Conclusion

Considering the currently unmet medical need to effectively overcome and treat ABCB1-induced MDR and also acknowledging the fact that this efflux-pump also plays a putative dominant role in cancer stem cells renders the successful discovery of new strategies a top priority in cancer research. MDR-selective agents represent a promising compound class with the potential of resolving ABCB1-mediated MDR. To this end, this is the first study, which aims to model and elucidate the SAR of MDR-selective and non-MDR-selective compounds. It is shown, that sophisticated data preprocessing methods in conjunction with highly complex ML algorithms are required to yield predictive models. Thus, suggesting that the underlying SAR is highly non-linear and rather complex. However, the careful interpretation of the most important predictors suggested that 2D-autocorrelation vectors that resemble molecular mass are among the most useful means to categorize MDR-selective compounds. In conclusion, the presented Random Forests classification model represents a promising starting point for the *in-silico* characterization of MDR-selective compounds. Further exploration of these promising anticancer agents will hopefully bring up new candidates that harbor the potential to overcome ABCB1-induced chemotherapeutic resistance in a very short time.

References

1. Longley, D.B. and P.G. Johnston, *Molecular mechanisms of drug resistance.* J. Pathol., 2005. **205**(2): p. 275-92.
2. Szakacs, G., et al., *Targeting multidrug resistance in cancer.* Nature Reviews Drug Discovery, 2006. **5**(3): p. 219-34.
3. Linton, K.J. and C.F. Higgins, *Structure and function of ABC transporters: the ATP switch provides flexible control.* Pflugers Arch., 2007. **453**(5): p. 555-67.
4. Higgins, C.F. and K.J. Linton, *The ATP switch model for ABC transporters.* Nat Struct Mol Biol., 2004. **11**(10): p. 918-26.
5. Aller, S.G., et al., *Structure of P-glycoprotein reveals a molecular basis for poly-specific drug binding.* Science, 2009. **323**(6922): p. 1718-22.
6. Ecker, G.F. and P. Chiba, *Transporters as Drug Carriers: Structure, Function, Substrates*. Methods and Principles in Burgers Medicinal Chemistry, ed. R. Mannhold, H. Kubinyi, and G. Folkers. Vol. 44. 2009. 429.
7. Hegedus, C., et al., *Interaction of ABC multidrug transporters with anticancer protein kinase inhibitors: substrates and/or inhibitors?* Curr Cancer Drug Targets, 2009. **9**(3): p. 252-72.
8. Hegedus, C., et al., *Interaction of nilotinib, dasatinib and bosutinib with ABCB1 and ABCG2: implications for altered anti-cancer effects and pharmacological properties.* British Journal of Pharmacology, 2009. **158**(4): p. 1153-1164.
9. Türk, D. and G. Szakacs, *Relevance of multidrug resistance in the age of targeted therapy.* . Curr Opin Drug Discov Devel., 2009. **12**(2): p. 246-52.
10. Szakacs, G., et al., *The role of ABC transporters in drug absorption, distribution, metabolism, excretion and toxicity (ADME-Tox).* Drug Discov Today., 2008. **13**(9-10): p. 379-93.
11. Aänismaa, R., E. Gatlik-Landwojtowicz, and A. Seelig, *P-glycoprotein senses its substrates and the lateral membrane packing density: consequences for the catalytic cycle.* Biochemistry, 2008. **47**(38): p. 10197-207.
12. Raviv, Y., et al., *Photosensitized labeling of a functional multidrug transporter in living drug-resistant tumor cells.* J Biol Chem, 1990. **265**(7): p. 3975-80.

13. Pleban, K. and G.F. Ecker, *Inhibitors of p-glycoprotein--lead identification and optimisation.* Mini Reviews in Medicinal Chemistry, 2005. **5**(2): p. 153-163.
14. Pleban, K., et al., *Targeting drug-efflux pumps -- a pharmacoinformatic approach.* Acta Biochim Pol., 2005. **52**(3): p. 737-40.
15. Palmeira, A., et al., *New uses for old drugs: pharmacophore-based screening for the discovery of Pglycoprotein inhibitors.* Chem Biol Drug Des, 2011. **78**(1): p. 57-72.
16. Colabufo, N.A., et al., *Perspectives of P-Glycoprotein Modulating Agents in Oncology and Neurodegenerative Diseases: Pharmaceutical, Biological, and Diagnostic Potentials.* J Med Chem, 2010. **53**: p. 1883-1897.
17. Nobili, S., et al., *Pharmacological strategies for overcoming multidrug resistance.* Curr Drug Targets, 2006. **7**(7): p. 861-79.
18. Wu, C.P., C.H. Hsieh, and Y.S. Wu, *The Emergence of Drug Transporter-Mediated Multidrug Resistance to Cancer Chemotherapy.* Molecular Pharmaceutics, 2011. **26**(Adv Onl).
19. Hall, M.D., M.D. Handley, and M.M. Gottesman, *Is resistance useless? Multidrug resistance and collateral sensitivity.* Trends in Pharmacological Sciences, 2009. **30**(10): p. 546-556.
20. Hall, M.D., et al., *Synthesis and structure-activity evaluation of isatin-ß-thiosemicarbazones with improved selective activity toward multidrug-resistant cells expressing P-glycoprotein.* Journal of Medicinal Chemistry, 2011. **54**(16): p. 5878-5889.
21. Türk, D., et al., *Identification of compounds selectively killing multidrug-resistant cancer cells.* Cancer Research, 2009. **69**(21): p. 8293-8301.
22. Warr, J.R., A. MBamford, and D.M. Quinn, *The preferential induction of apoptosis in multidrug-resistant KB cells by 5-fluorouracil.* Cancer Lett., 2002. **175**(1): p. 39-44.
23. Nicholson, K.M., et al., *Preferential killing of multidrug-resistant KB cells by inhibitors of glucosylceramide synthase.* Br. J. Cancer, 1999. **81**(3): p. 423-30.
24. Nicholson, K.M., et al., *LY294002, an inhibitor of phosphatidylinositol-3-kinase, causes preferential induction of apoptosis in human multidrug resistant cells.* Cancer Letters, 2003. **190**(1): p. 31-6.
25. Szakacs, G., et al., *Predicting drug sensitivity and resistance: profiling ABC transporter genes in cancer cells.* Cancer Cell, 2004. **6**(2): p. 129-37.
26. Hall, M.D., et al., *Synthesis, activity, and pharmacophore development for isatin-beta-thiosemicarbazones with selective activity toward multidrug-resistant cells.* J Med Chem, 2009. **52**(10): p. 3191-204.
27. Ludwig, J.A., et al., *Selective toxicity of NSC73306 in MDR1-positive cells as a new strategy to circumvent multidrug resistance in cancer.* Cancer Research, 2006. **66**: p. 4808-4815.
28. Wu, C.P., et al., *Evidence for dual mode of action of a thiosemicarbazone, NSC73306: a potent substrate of the multidrug resistance–linked ABCG2 transporter.* Mol. Cancer Ther., 2007. **6**: p. 3287-3296.
29. Cheng, A. and S.L. Dixon, *In silico models for the prediction of dosedependent human hepatotoxicity.* J. Comput.-Aided Mol. Des., 2003. **17**: p. 811-823.
30. Digles, D. and G.F. Ecker, *Self-Organizing Maps for In Silico Screening and Data Visualization.* Molecular Informatics, 2011: p. Adv. Onl.
31. Dixon, S.L. and H.O. Villar, *Investigation of classification methods for the prediction of activity in diverse chemical libraries.* J. Comput.-Aided Mol. Des., 1999. **13**: p. 533-545.
32. Ivanciuc, O., *Applications of Support Vector Machines in Chemistry*, in *Reviews in Computational Chemistry*, K. Lipkowitz and T.R. Cundari, Editors. 2007, Wiley-VCH: Weinheim. p. 291-400.
33. Susnow, R.G. and S.L. Dixon, *Use of robust classification techniques for the prediction of human cytochrome P450 2D6 inhibition.* J. Chem. Inf. Comput. Sci., 2003. **43**: p. 1308-1315.
34. Svetnik, V., et al., *Application of Breiman's random Forests to modeling structure-activity relationships of pharmaceutical molecules*, in *Multiple Classifier Systems*, F. Roli, J. Kittler, and T. Windeatt, Editors. 2004, Springer-Verlag: Berlin. p. 334-343.
35. Demel, M.A., *Ensemble Rule-Based Classification of Substrates of the Human ABC-Transporter ABCB1 Using Simple Physicochemical Descriptors.* Molecular Informatics, 2010. **29**: p. 233-242.

36. Demel, M.A., et al., *Comparison of Contemporary Feature Selection Algorithms: Application to the Classification of ABC-Transporter Substrates.* QSYR Comb. Sci., 2009. **28**(10): p. 1087-1091.
37. Schwaha, R. and G.F. Ecker, *Use of shape similarities for the classification of P-glycoprotein substrates and nonsubstrates.* Future Med Chem, 2011. **3**(9): p. 1117-28.
38. de Cerqueira Lima, P., et al., *Combinatorial QSAR modeling of P-glycoprotein substrates.* J Chem Inf Model, 2006. **46**(3): p. 1245-54.
39. Huang, J., et al., *Identifying P-glycoprotein substrates using a support vector machine optimized by a particle swarm.* J Chem Inf Model, 2007. **47**(4): p. 1638-47.
40. Klepsch, F., P. Chiba, and G.F. Ecker, *Exhaustive Sampling of Docking Poses Reveals Binding Hypotheses for Propafenone Type Inhibitors of P-Glycoprotein.* PLoS Comput Biol., 2011. **7**(5): p. e1002036.
41. Klepsch, F., et al., *Pharmacoinformatic approaches to design natural product type ligands of ABC-transporters.* Current Pharmaceutical Design, 2010. **16**(15): p. 1742-1752.
42. Chen, L., et al., *ADME evaluation in drug discovery. 10. Predictions of P-glycoprotein inhibitors using recursive partitioning and naive Bayesian classification techniques.* Molecular Pharmaceutics, 2011. **8**(3): p. 889-900.
43. Demel, M.A., et al., *Predicting ligand interactions with ABC transporters in ADME.* Chemistry & Biodiversity, 2009. **6**(11): p. 1960-1969.
44. Ecker, G.F., et al., *The Importance of a Nitrogen Atom in Modulators of Multidrug Resistance.* Molecular Pharmacology, 1999. **56**: p. 791-796.
45. Hong, H., Xie, Q., Ge, W., Qian, F., Fang, F., Shi, L., Su, Z., Perkins, R., and Tong, W., *Mold2, molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics.* Journal of Chemical Information and Modeling, 2008. **48**(7): p. 1337-1344.
46. FDA, F.a.D.A. *Mold2 News and Publications*. [cited 2011 2011-11-04]; Available from: http://www.fda.gov/ScienceResearch/BioinformaticsTools/Mold2/ucm144528.htm.
47. Willett, P., *Dissimilarity-Based Algorithms for Selecting Structurally Diverse Sets of Compounds.* Journal of Computational Biology, 1999. **6**(3-4): p. 447-57.
48. Whitley, D.C., M.G. Ford, and D.J. Livingstone, *Unsupervised Forward Selection: A Method for Eliminating Redundant Variables.* Journal of Chemical Informationa and Computational Sciences, 2000. **40**: p. 1160-1168.
49. Breiman, L., *Random Forestss.* Machine Learning, 2001. **45**(1): p. 5-32.
50. Friedman, J.H., *Greedy Function Approximation: A Gradient Boosting Machine*. 2000, IMS Reitz Lecture.
51. Friedman, J.H., *Stochastic Gradient Boosting*, in *Technical Discussion of TreeNEt*. 1999.
52. Burges, C.J.C., *A Tutorial on Support Vector Machines for Pattern Recognition.* Data Mining and Knowledge Discovery, 1998. **2**: p. 121-167.
53. Cortes, C. and V. Vapnik, *Support-Vector Networks.* Machine Learning, 1995. **20**.
54. Kearns, M., *Thoughts on hypothesis boosting*. 1988.
55. Witten, I.H. and E. Frank, *Data Mining - Practical Machine Learning Tools and Techniques*. 2005: Elsevier. 558.
56. Kotsianis, S.B., *Supervised Machine Learning: A Review of Classification Techniques.* Informatica, 2007. **31**: p. 249-268.
57. Hastie, T., R. Tibshirani, and J.H. Friedman, *The Elements of Statistical Learning - Data Mining, Inference and PRediction*. 2008.
58. Sutton, C.D., *Classification and Regression Trees, Bagging, and Boosting*, in *Handbook of Statistics-Data Mining and Data Visualization* 2005, Elseier. p. 303-329.
59. Campbell, C., *Kernel methods: a survey of current techniques.* Neurocomputing, 2002. **48**: p. 63-84.
60. Sanchez, V.D., *Advanced support vector machines and kernel methods.* Neurocomputing, 2003. **55**: p. 5-20.
61. Breiman, L., et al., *CART: Classification and Regression Trees*. 1983, Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

62. Freund, Y. and R.E. Schapire, *A decision-theoretic generalization of online learning and an application to boosting.* J. Comput. System Sci., 1997. **55**(119-139).
63. He, P., et al., *Improving the classification accuracy in chemistry via boosting technique.* Chemom. Intell. Lab. Syst, 2004. **70**: p. 39-46.
64. Friedman, J.H., *Greedy Function Approximation: A Gradient Boosting Machine.* Ann. Stat., 2001. **29**: p. 1189-1202.
65. Ridgeway, G., *The gbm package*.
66. Veropoulos, K., C. Campbell, and N. Cristianini. *Controlling the Sensitivity of Support Vector Machines*. in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI99)*. 1999.
67. Smola, A.J. and B. Schölkopf, *On a kernel-based method for pattern recognition, regression, approximation, and operator inversion.* Algorithmica, 1998. **22**: p. 211-231.
68. Schölkopf, B., et al., *Input space versus feature space in kernel-based methods.* IEEE Trans. Neural Netw., 1999. **10**: p. 1000-1017.
69. Press, W., et al., *Numerical Recipes: The Art of Scientific Computing* 3rd ed. 2007, New York: Cambridge Univeristy Press.
70. Karatzoglou, A., D. Meyer, and K. Hornik, *Support Vector Machines in R.* Journal of Statistical Software, 2006. **15**(9).
71. Matthews, B.W., *Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozime.* Biochim. Biophys. Acta, 1975. **405**: p. 442-451.
72. Benigni, R., *Structure-activity Relationship Studies of Chemical Mutagens and Carcinogens: Mechanistic Investigations and Prediction Approaches.* Chem. Rev., 2005. **105**: p. 1767-1800.
73. Hillebrecht, A. and G. Klebe, *Use of 3D QSAR models for database screening: a feasibility study.* J. Chem. Inf. Model, 2008. **48**: p. 384-396.
74. Swets, J.A., R.M. Dawes, and J. Monahan, *Better Decisions Through Science.* Sci. Am., 2000. **283**: p. 82-87.
75. Rizzi, A. and A. Fioni, *Virtual Screening using PLS Discriminant Analysis and ROC Curve Approach: An Application Study on PDE4 Inhibitors.* J Chem Inf Model, 2008. **41**: p. 1686-1692.

Author Contribution

I collected the chemical information of this chapter from the cited reference (see main text). I designed the experiments. I wrote all the R-scripts. Mag. Waglechner provided substantial advice for the implementation of UFS in R (specifically, he had the idea of the "while-loop" instead of my initial "nested-for" variant prior to step 5 of this algorithm). I performed all the calculations and evaluated the results. I wrote this manuscript.

Supplementary Information

LDA and DT:

Supp Table ST 1: LDA and DT model performance in 10-fold CV using 11 simple and intuitive ADME parameters.

Model	Sens	Spec	ROC
LDA	0.654	0.724	0.612
DT	0.621	0.543	0.578

Model Training and parameter tuning:

Supp Table ST 2: Training performance of RF models across the tuning parameter mtry; Validation was performed using three independent runs of 10-fold CV. The "best model" was selected on basis of ROC. ROC denotes the "area under the ROC curve" using the trapezoidal rule. mtry denotes the number of predictors used for the generation of each tree; the number of trees was held constant with n.tree=500.

mtry	Sens	Spec	ROC	Sens SD	Spec SD	ROC SD	Selected
2	0.906	0.764	0.882	0.0565	0.085	0.0501	
16	0.919	0.743	0.888	0.0534	0.0987	0.048	*
30	0.89	0.756	0.875	0.0662	0.0869	0.0498	

Supp Table S 3: Training performance of GBM models across the two tuning parameters interaction.depth and n.trees; Validation was performed using three independent runs of 10-fold CV. The "best model" was selected on basis of ROC. ROC denotes the "area under the ROC curve" using the trapezoidal rule. Interaction.depth limits the tree size and n.trees denotes the number of trees that are used to establish the boosted ensemble; the shrinkage parameter was held constant at a value of s=0.1.

interaction.depth	n.trees	Sens	Spec	ROC	Sens SD	Spec SD	ROC SD	Selected
1	**50**	0.896	0.675	0.851	0.0781	0.107	0.0665	
1	**100**	0.882	0.702	0.854	0.0797	0.0998	0.0626	
1	**150**	0.876	0.704	0.853	0.0785	0.0996	0.0611	
2	**50**	0.881	0.708	0.857	0.0989	0.108	0.0623	
2	**100**	0.873	0.725	0.858	0.0894	0.106	0.0652	
2	**150**	0.88	0.731	0.855	0.0715	0.104	0.0639	
3	**50**	0.876	0.708	0.86	0.0907	0.112	0.0696	
3	**100**	0.868	0.731	0.86	0.0777	0.0986	0.063	
3	**150**	0.874	0.733	0.862	0.0784	0.104	0.0601	*

Supp Table ST 4: Training performance of SVM models across the three tuning parameters degree, C and scale; Validation was performed using three independent runs of 10-fold CV. The "best model" was selected on basis of ROC. ROC denotes the "area under the ROC curve" using the trapezoidal rule. The parameter "degree" denotes the degree of the polynomial function deployed by the kernel. C is the cost parameter which defines the constraints violation.

degree	C	scale	Sens	Spec	ROC	Sens SD	Spec SD	ROC SD	Selected
1	**0.1**	**0.001**	1	0	0.837	0	0	0.061	
1	**0.1**	**0.01**	0.984	0.468	0.841	0.0266	0.125	0.0594	
1	**0.1**	**0.1**	0.915	0.7	0.862	0.0598	0.11	0.0563	
2	**0.1**	**0.001**	1	0	0.837	0	0	0.0604	
2	**0.1**	**0.01**	0.938	0.607	0.854	0.0553	0.124	0.0531	
2	**0.1**	**0.1**	0.887	0.725	0.868	0.0763	0.104	0.048	
3	**0.1**	**0.001**	1	0.0145	0.837	0	0.0267	0.0602	
3	**0.1**	**0.01**	0.931	0.64	0.858	0.0529	0.118	0.0492	
3	**0.1**	**0.1**	0.855	0.762	0.861	0.0752	0.0799	0.0415	
1	**1**	**0.001**	0.984	0.468	0.841	0.0266	0.125	0.0594	
1	**1**	**0.01**	0.915	0.7	0.862	0.0598	0.11	0.0563	
1	**1**	**0.1**	0.867	0.719	0.863	0.0734	0.111	0.0613	
2	**1**	**0.001**	0.925	0.611	0.849	0.0559	0.129	0.0554	
2	**1**	**0.01**	0.893	0.727	0.872	0.0638	0.0988	0.0491	*

2	**1**	**0.1**	0.839	0.774	0.848	0.0752	0.102	0.0441
3	**1**	**0.001**	0.917	0.652	0.855	0.056	0.131	0.0538
3	**1**	**0.01**	0.88	0.741	0.871	0.0764	0.108	0.0516
3	**1**	**0.1**	0.831	0.768	0.847	0.0785	0.0828	0.0436
1	**10**	**0.001**	0.915	0.7	0.862	0.0598	0.11	0.0563
1	**10**	**0.01**	0.867	0.719	0.863	0.0734	0.111	0.0613
1	**10**	**0.1**	0.845	0.719	0.852	0.075	0.115	0.0656
2	**10**	**0.001**	0.893	0.722	0.862	0.0696	0.104	0.0538
2	**10**	**0.01**	0.874	0.741	0.864	0.0706	0.0899	0.0489
2	**10**	**0.1**	0.807	0.774	0.845	0.0772	0.0922	0.0428
3	**10**	**0.001**	0.888	0.718	0.866	0.0622	0.0968	0.055
3	**10**	**0.01**	0.858	0.762	0.859	0.0829	0.082	0.0442
3	**10**	**0.1**	0.831	0.768	0.847	0.0785	0.0828	0.0436

Source Code of the UFS algorithm in R:

```
#NOTE:
#the ufs function expects three parameters:1-the descriptor matrix;2-the
predefined value for rsqmax; 3-the predefined value #  for eps;
# the function returns the ufs selected descriptor matrix; additional
information is printed out onto the console

ufs=function(x,rsqmax=0.75,eps=0.05){
tmp=vector()
minimas=vector()
torem=vector()
todisc=vector()

  #data is the orignal descriptor matrix, that is kept and finally
  #returned by this function with only the selected descriptors
  data=x
  #1. mean-center the columns of x

x.cent=scale(x,center=T,scale=F)
x.cent=as.data.frame(x.cent)
                x.sd=sd(x,na.rm=T)
  #2.reject columns with length < eps
      torem=which(x.sd<eps)
  if(length(torem)>=1){
      x.cent=x.cent[,-torem]
            x.sd=x.sd[-torem]
  }
  if(dim(x.cent)[2]==0){
      print("EPS is too big!!!!")
            return(-1)
}
  if(dim(x.cent)[2]<=2){
   print("EARLY ABORT: after applying EPS-criterion less than 2 cols
remained")
      give=c(colnames(x.cent))
      data=data[,give]
   return(data)
}

#3.normalize to unit length
    for(j in 1: dim(x.cent)[2]){
    x.cent[,j]=x.cent[,j]/x.sd[j]
```

```
  }

#4.calculate the correlation matrix
  cor.x.cent=cor(x.cent)^2
    first_col=floor((which.min(cor.x.cent)-1)/ncol(cor.x.cent))+1
      first_row=which.min(cor.x.cent)-(first_col-1)*ncol(cor.x.cent)
      #tokeep is a string with the names of the descriptors that shall be
      kept

tokeep.names=c(rownames(cor.x.cent)[first_row],colnames(cor.x.cent)[first_c
ol])
  #sel=data.frame to start
    sel=as.data.frame(cbind(x.cent[,first_row],x.cent[,first_col]))
      tmp.mat=as.matrix(x.cent[,c(-first_row,-first_col)])
             todisc1=which(cor(sel[,1],tmp.mat)^2>=rsqmax)
              todisc2=which(cor(sel[,2],tmp.mat)^2>=rsqmax)
                tmp=0
                  todisc=c(todisc1,todisc2)
                     todisc=sort(todisc[!duplicated(todisc)])
             if(!length(todisc)==0){
              tmp.mat=as.matrix(tmp.mat[,-todisc])
             }
    i=1
    while(dim(tmp.mat)[2]>0){

#5.orthogonalize the two selected descriptors
      qr=qr(sel,LAPACK=F)
        qr=qr.Q(qr)

#6. for each remaining column calculate its squared correlation coefficient
to the selected columns
                   tokeep.values=apply(cor(qr,tmp.mat)^2,2,sum)
                    tokeep=which(tokeep.values<rsqmax)
                      tokeep.values=tokeep.values[tokeep]
                          tmp.mat=as.matrix(tmp.mat[,tokeep])

                        if(dim(tmp.mat)[2]==0)
                          break
                     min.index=which.min(tokeep.values)
                        tokeep.names=c(tokeep.names,names(min.index))
                              sel=cbind(sel,tmp.mat[,min.index])
                                    tmp.mat=as.matrix(tmp.mat[,-min.index])
i=i+1
     }
     data=data[,tokeep.names]
return(data)
}
```

Chapter 16: Manuscript IV - Retrospective Analysis of Structure-Selectivity Relationships of Compounds Selectively Killing ABCB1-overexpresing Cells using Network-like Similarity Graphs

Abstract

The upregulation of multidrug efflux pumps of the ABC-transporter family, such as ABCB1 (P-glycoprotein, P-gp), has been identified as one of the main contributors to the development of multidrug resistance (MDR) in various haematological but also solid malignancies and hence is often associated with a poor chemotherapeutic treatment response, which finally severely compromises clinical survival in cancer patients. For more than three decades the main research efforts concentrated to overcome cancer associated MDR via inhibition of these ABC-transporters. Despite promising preclinical results, the concept of direct transporter inhibition could so far not be realized in clinical practice. Therefore, alternative strategies need to be urgently developed to resolve clinical MDR. Recently, a novel class of isatin-β-thiosemicarbazones (IBTs) was identified, that preferentially killed ABC-transporter overexpressing MDR cells more efficiently than non-resistant parental cells, which do not express the respective transporter. These compounds were termed "MDR-selective molecules" or "collateral sensitive" compounds. In this study, the structure-selectivity relationships (SSR) of a data set of 41 MDR-selective IBTs is explored by means of network-like similarity graphs (NSGs) implemented in the open-source SARANEA software package. This allows a detailed quantitative description and categorization of the IBT SSR. Furthermore, by analysing local SSR environments compound selectivity determinants as well as so-called "selectivity cliff markers" are identified. Additionally, the use of so-called "marker sets" enables the identification of selective compounds having structural features that render them as potential toxic. Overall, this retrospective analysis provides useful information on the chemical nature of MDR-selective cytotoxicity of IBTs and might serve as fruitful input for further exploratory lead-optimization efforts regarding MDR-selective molecules.

Introduction

Cytototoxic or cytostatic pharmacotherapy represents the current gold standard for the treatment of many different types of cancer. However, the occurrence of acquired or intrinsic drug resistance is responsible for approximately 90% of treatment failure and is associated with poor overall survival (Longley & Johnston 2005). Multidrug resistance (MDR), which is defined as the simultaneous resistance to pharmacologically, mechanistically and structurally unrelated drugs, is often mediated by the overexpression of members of the human multidrug ABC-transporter family (Szakács et al. 2006; Sarkadi & Szakács 2010; Klepsch, Stockner, et al. 2010; Klepsch, Jabeen, et al. 2010). These transmembrane proteins reduce intracellular drug concentrations below therapeutic concentrations by actively exporting a wide range of hydrophobic and also amphiphilic molecules out of living cells irrespective of the concentration gradient. Besides, this therapy limiting role in cancer, ABC-transporters confer also physiological functions. They are endogenously expressed in the liver, the kidney, the intestines and the blood-brain barrier (Giacomini et al. 2010). There these proteins are essential for physiological homeostasis, while at the same time interfere with drug absorption,

distribution and elimination. Thereby giving rise to potential drug side effects or drug-drug interactions.

Up to date, at least 15 ABC-transporters have been linked to MDR *in vitro* and three prominent members (ABCB1, ABCC1, ABCG2) have been convincingly associated with clinical features of cancer drug resistance. Of these, ABCB1 represents the best characterized member and a variety of antineoplastic agents commonly used in the clinic have been shown to be transported by ABCB1. These include structurally distinct drugs such as paclitaxel, vinblastine and daunorubicin. Additionally, most of the novel rationally developed tyrosine kinase inhibitors are also substrates of ABCB1 (Shukla et al. 2012).

So far a plethora of different strategies to overcome or circumvent ABCB1-conferred MDR have been thoroughly investigated, including the design of inhibitors that block transport capacity and restore intracellular anticancer drug concentrations, the design of novel molecules that evade molecular recognition by the transporter and thereby "escape" efflux, and very recently the discovery of MDR-selective small molecules that display a selectively lethal phenotype to ABCB1-overexpressing cells.

These MDR-selective molecules, which are also termed "collateral sensitive" compounds have attracted a lot of attention in the recent years (Goldsborough et al. 2011; Nobili et al. 2011; Hall, Handley, et al. 2009). These molecules reverse the mechanism of MDR into a weakness. Their cytotoxicity is potentiated in cells that are overexpressing ABCB1 rather than reduced. In other words these agents are selective for cells expressing ABCB1. In in-vitro assays MDR-selectivity is usually assessed by calculating the ratio of the IC_{50} value of a compound under investigation in a parental cancer cell line divided by its IC_{50} value in the corresponding MDR cell line (selectivity ratio, SR) (Türk et al. 2009). The parental cancer cell line does not express the ABC-transporter. In recent years, many MDR-selective molecules have been identified and the further exploration of this novel pharmacological class is of high medical relevance (Gillet & Michael M Gottesman 2012; Pluchino et al. 2012). This class of molecules harbours the potential to prevent the occurrence of MDR during standard anticancer chemotherapy when administered in an adjuvant setting. Alternatively, these agents might also be able to resensitize already treatment-refractory cancers through their selective killing of MDR cells within the heterogeneous tumour cell population.

Network-like similarity graphs (NSGs) are simple and very intuitive tools for the systematic comparison and also for the visualization of potency or selectivity distributions and similarity relationships within a pharmacological class (Wawer et al. 2008; Wawer & Bajorath 2009; Wawer et al. 2009; Lounkine et al. 2010; Peltason, Hu, et al. 2009; Peltason, Weskamp, et al. 2009). In this type of graphs molecules are displayed as nodes and edges, which connect these nodes, highlight the similarity between two or more nodes. For the similarity calculation the MACCS-based Tanimoto coefficient is used. Nodes are color-coded according to their activity/selectivity values. The size of the node describes the contribution of a given molecule to a certain SAR type. In summary, NSGs facilitate the simultaneous investigation of similarity networks and multiple SAR components. It is the primary objective of this contribution to apply NSGs (which can be calculated with the open-source software

SARANEA (Lounkine et al. 2010)) to a compound set of MDR-selective molecules in order to determine the structural determinants for the selective cell killing properties of this pharmacological class. In its original implementation, selectivity NSGs are intended to compare a dataset with measured activities for one target to measured activities to another target. Here, we adapt this original implementation, in that respect, that we compare the measured IC_{50} values of a ABCB1-negative to a ABCB1-expressing cell line; rather than using data from two different (protein) targets.

Material and Methods

Data Sets

Selectivity Data Set: The 41 chemical structures used in this study were compiled from two literature references (Hall, Salam, et al. 2009; Hall et al. 2011). All 41 molecules were originally synthesized and pharmacologically evaluated in the group of Michael M. Gottesman at the National Cancer Institute. Eleven molecules with measured cellular toxicity values were taken from Hall, 2009 and the remaining 30 molecular structures were taken from Hall, 2011. The assessment of in-vitro cytotoxicity and selectivity values for these molecules is described in detail by Hall et al. (Hall, Salam, et al. 2009; Hall et al. 2011). Briefly, the MTT assay was used to determine the cytotoxicity in the parental ABCB1-negative KB-3-1 cell line. KB-3-1 cells are human adenocarcinoma cells. Additionally, the molecules were also evaluated in the KB-V1 cell line. This human ABCB1 overexpressing adenocarcinoma cell line was developed by stepwise selection of KB-3-1 cells in the presence of the ABCB1 substrate vinblastine (Shen et al. 1986). Finally, MDR-selectivity was quantified by calculating the ratio of IC_{50} against KB-3-1 parental cells divided by its IC_{50} against KB-V1 cells. A selectivity ratio (SR) >1 denotes a MDR-selective molecule, since it kills ABCB1-overexpressing cells more effectively than the parental cell line (Hall, Handley, et al. 2009). The chemistry of this data set is described in the result section and the chemical structures are provided in the supplementary information.

Marker Sets: Furthermore, marker sets that are distributed alongside with the SARANEA software are also employed in this study. These marker sets can be added to the calculated NSG and help to identify active compounds having structural features that might render them problematic (Lounkine et al. 2010). In case these marker compounds show a molecular similarity above a predefined similarity threshold, they are connected to the respective active molecules via edges in the NSG. In this contribution, organ-focussed (hematopoietic system, liver and kidney) toxicity sets were used.

Molecular Structure Representation

Chemical structures were represented as SMILES strings and as binary MACCS fingerprints. For the calculation of the publicly available 166-bit MACCS fingerprint the RDKit implemented in KNIME 2.2.3 was used. The chemical structures as well as biological data are provided in the supplementary information section enclosed to this contribution.

SARANEA Software for SARI Determination and NSG modelling

The SARANEA software which is used to calculate NSGs and to derive the SARIs for the data set under investigation is distributed free-of-charge under the GNU General Public License and was obtained from: www.lifescienceinformatics.uni-bonn.de. It is distributed together with three public Java libraries. These are the Java Universal Network/Graph Framework (JUNG2), JChemPaint, and JOELib.

Assessment of Molecular Similarity: In SARANEA the chemical structure of molecules is compared using a molecular fingerprint representation. In this study the MACCS 166-bit fingerprint is used as input for similarity assessment. As similarity metric, SARANEA utilizes the Tanimoto coefficient (Tc), which is calculated as follows for two molecules A and B:

$$Tc = \frac{c}{a + b - c}$$

From this formula, *a* and *b* correspond to the number of bits present in each molecule (A or B) and *c* corresponds to the number of bits present in both molecules (A and B).

Derivation of the SAR Index: In order to capture complement SAR types the SARI function consists of two sub-scores (Peltason & Bajorath 2007; Peltason, Hu, et al. 2009; Peltason, Weskamp, et al. 2009). These are the continuity score ($score_{cont}$) and the discontinuity score ($score_{disc}$). Prior to the calculation of these two scores, "raw" scores are derived. The $score_{cont}$ represents the potency-weighted mean of the reciprocal similarity values of a data set:

$$cont_{raw} = \frac{\sum_{\{(i,j)|i \neq j\}} (weight(i,j) * \frac{1}{(1 + sim(i,j))})}{\sum_{\{(i,j)|i \neq j\}} weight(i,j)}$$

where the weight is defined as,

$$weight(i,j) = \frac{pot_i * pot_j}{1 + |pot_i - pot_j|}$$

In this formula *sim(i,j)* denotes the pairwise compound similarity between the molecules *i* and *j* and pot_i and pot_j represent the log-scaled activity values. From this it becomes apparent that dissimilar compounds with comparable and high activities will contribute most to the score.

Compounds displaying a discontinuous SAR landscape (similar molecules showing large differences in potency/selectivity) are captured by the second sub-score, $score_{disc}$, which is derived using the mean potency difference of all compound pairs that exceed a predefined similarity threshold (e.g.: MACCS Tc = 0.65) and potency threshold ("1" on a log-scale, i.e. ten-fold difference):

$$disc_{raw} = mean(|pot_i - pot_j| * sim(i,j))$$

where *mean* is defined as:

$$mean = \{(i,j)|sim(i,j) > tsh, |pot_i - pot_j > 1|\}$$

These two "raw" scores are subsequently subjected to *z*-standardisation. For this, the mean and the SD of a reference panel (compiled from BindingDB; see SARANEA manual for details) are used. The final $score_{cont}$ and $score_{disc}$ values are then derived via mapping these z-scores on the range [0,1] by calculating the cumulative probability for each score assuming a normal distribution. The final SARI is calculated from these two complementary measures according to (Peltason & Bajorath 2007):

$$SARI = \frac{1}{2}(score_{cont} + (1 - score_{disc}))$$

For calculations utilizing selectivity sets, the potencies are exchanged with the corresponding selectivity values (Lounkine et al. 2010).

Compound discontinuity score: The $score_{disc}$ component of the SARI refers to the overall discontinuity of the data set under investigation. In order to estimate the individual contribution of single compounds to this score, a modified compound-specific $score_{disc}$ has been introduced (Wawer et al. 2008). The main differences compared to the overall $score_{disc}$ are that potency contributions are not considered and that only the nearest neighbors are used in the calculation. Selectivity-based scores for each compound are derived by replacing potency by selectivity ratios. Again, this score is normalized to the range [0,1]. Higher values denote potential activity/selectivity cliffs.

NSGs: The SARANEA software transfers similarity values that are calculated from the fingerprint representation of the molecules into a visual graph structure. These graphs consist of nodes and edges. Each node represents an individual molecule and nodes are connected via edges if their pairwise similarity exceeds the user-defined threshold. Usually edges are directed from nodes with lower potency/selectivity to nodes with higher potency/selectivity. In case of no potency/selectivity difference, edges remain undirected. A continuous colour gradient (green=low potency/selectivity; yellow=medium potency/selectivity, red=high potency/selectivity) is used to graphically denote the biological activity/selectivity of the molecules. The overall layout of the network is determined using the Fruchtermann-Rheingold force-directed algorithm, which visually clusters densely connected groups (Fruchterman & Reingold 1991; Wawer et al. 2008). In selectivity NSGs, green nodes reflect molecules selective for one target, whereas red nodes reflect molecules selective for the other target. Hence, in the context of this contribution green nodes reflect molecules more selective in parental KB-3-1 cells and red nodes denote molcules selective for ABCB1-expresing KB-V1 cells. Additionally, the size of each node represents the calculated compound $score_{disc}$. Hence, the compound-specific $score_{disc}$ and the node size convey the same information: the higher the score/the larger the node size, the more likely that the compound represents an activity/selectivity cliff marker.

Calculation of SAR pathways: SAR pathways are sequences of pair-wise compounds with chemical similarity identified in an NSG that are based on changes in activity along the calculated path (see supplementary information in (Lounkine et al. 2010) for details). In order to calculate a SAR pathway two criteria need to be defined: (i) two connected molecules have

to show an increase in potency. Compounds with identical potency are omitted from the pathway. (ii) every path is assigned a cost-function and the lowest cost between pairs of molecules (nodes) is finally selected. The cost for an edge between compounds *i* and *j* (connection between two nodes) is calculated as follows:

$$cost(edge_{ij}) = \frac{(1 + potdiff(i,j)^2)}{sim(i,j)}$$

In order to calculate the path with lowest overall cost, the Djikstra algorithm is applied (Djikstra 1959). The cost for an entire SAR pathway enclosing molecules *1,…,n* is simple the some of the individual costs of the edges

$$cost(path_{1,n}) = \sum_{i=1}^{n=l} cost(edge_{i,j+1})$$

This returns the shortest path between every connected pair of molecules. In order to evaluate different pathways between different molecule pairs, the following scoring function is applied

$$score(path_{i,j}) = \frac{potdiff(i,j)}{1 + deviation(path_{i,j})}\ x\ (\left|edges_{path_{i,j}}\right| - 1)$$

The formula for the score shown above considers the total potency difference as well as the smoothness of the potency gradient along the path (Wawer et al. 2009).

Results and Discussion

Characterization of the MDR-selectivity Data Set

Prior to any *in-silico* calculations the chemical composition of the 41 molecules under consideration was manually investigated (see Figure 1). The common structural denominator which is present in all instances is the nitrogen containing thiosemicarbazone, which is formally a condensation product between a thiosemicarbazide and the respective aldehyde or ketone. The majority of molecules (38/41) show a β-isatin substructure at position N1. Eleven of these also show a substitution or modification of the β-isatin scaffold. Additionally, out of the 41 molecules, 29 show an aromatic phenyl-substitution at N4. The majority of the N4-aromatic molecules also exhibits an additional *para*-substitution. Overall, 28 molecules show both substitutions; i.e. the N1-β-isatin as well as the N4 phenyl substitution. Among these, seven molecules exhibit a substituted/modified N1-β-isatin together with a para-substituted N4-phenyl moiety.

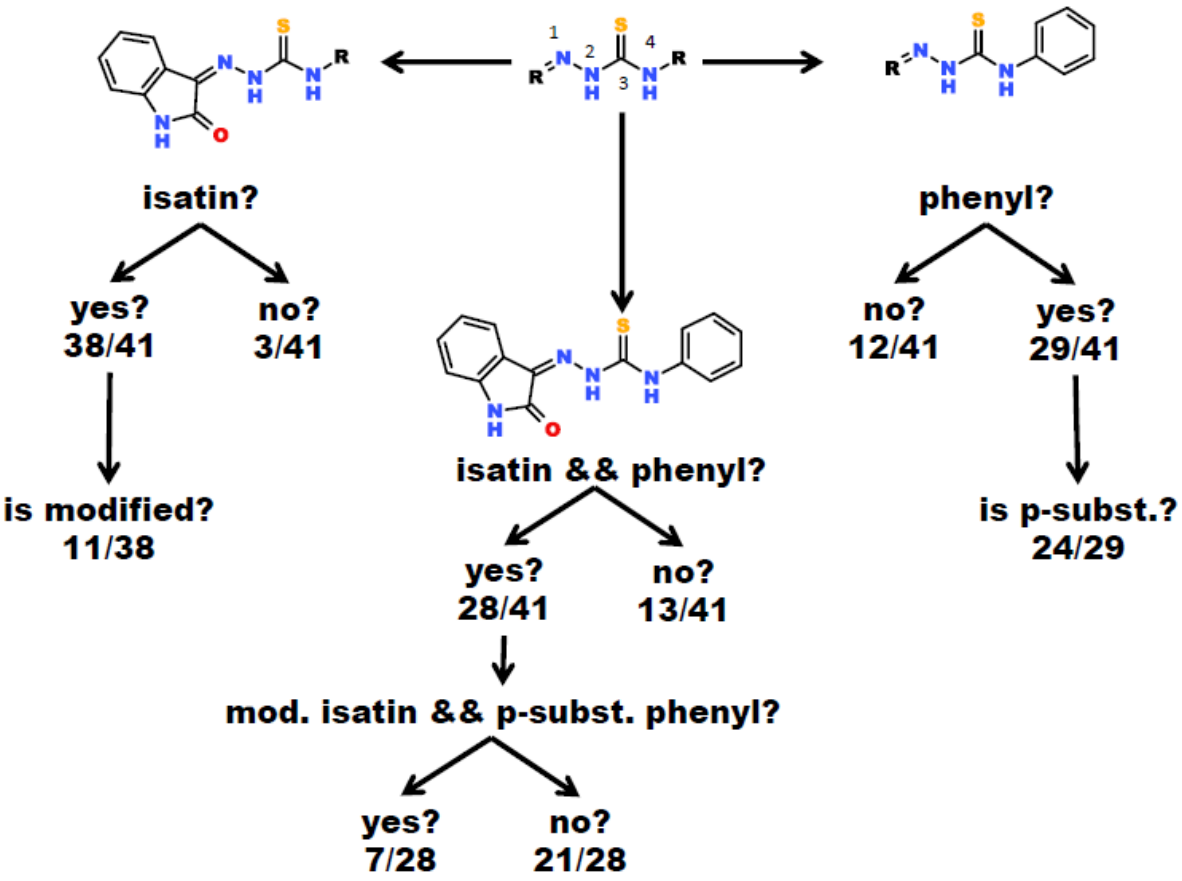


Figure 1: Chemical Composition of the data set.

Furthermore, the potencies as well as the physico-chemical properties of this data set were also investigated in a descriptive manner (see Table 1).

Table 1: Characteristics of the data set of 41 MDR-selective molecules

	min	max	mean	SD
Cell line & SR				
KB-3-1(IC_{50} nM)	1400	39300	13976.34	11427.28
KB-V1 (IC_{50} nM)	640	34300	6547.80	7853.84
SR	0.15	14.86	3.44	2.94
PhysChem property				
Weight	194	451	329.68	49.46
TPSA	79	141	95.00	16.17
SlogP	1	6	3.32	0.99
b_rotN	3	7	4.66	0.91
a_acc	2	5	2.27	0.59

Global SAR and SSR Features of MDR-selective molecules

NSGs are applied to provide a topological representation of similarity and potency/selectivity relationships within the MDR-selectivity data set under consideration. Figure 2A-C displays

the different appearances of the selectivity set (KB-3-1/KB-V1) at different similarity thresholds. The similarity between molecules was calculated using the Tanimoto coefficient (Tc) on basis of the publicly available 166-bit MACCS fingerprint. From Figure 2A-C it can be seen that all compounds (nodes) are connected via edges at Tc = 0.65. At higher Tc values (Figure 2C) several distinct sub-graphs or communities emerge. Some of these sub-graphs are constituted by molecules of equal selectivity values, whereas others are constituted of molecules with varying selectivity values (selectivity values are visualized by node color). Since the network topology of a NSG is solely determined by pairwise compound similarity, it is identical for potency and selectivity NSGs. Therefore, only the selectivity (KB-3-1/KB-V1) NSG is shown in Figure 2. Throughout this text the NSG derived using Tc = 0.65 will be used unless stated otherwise.

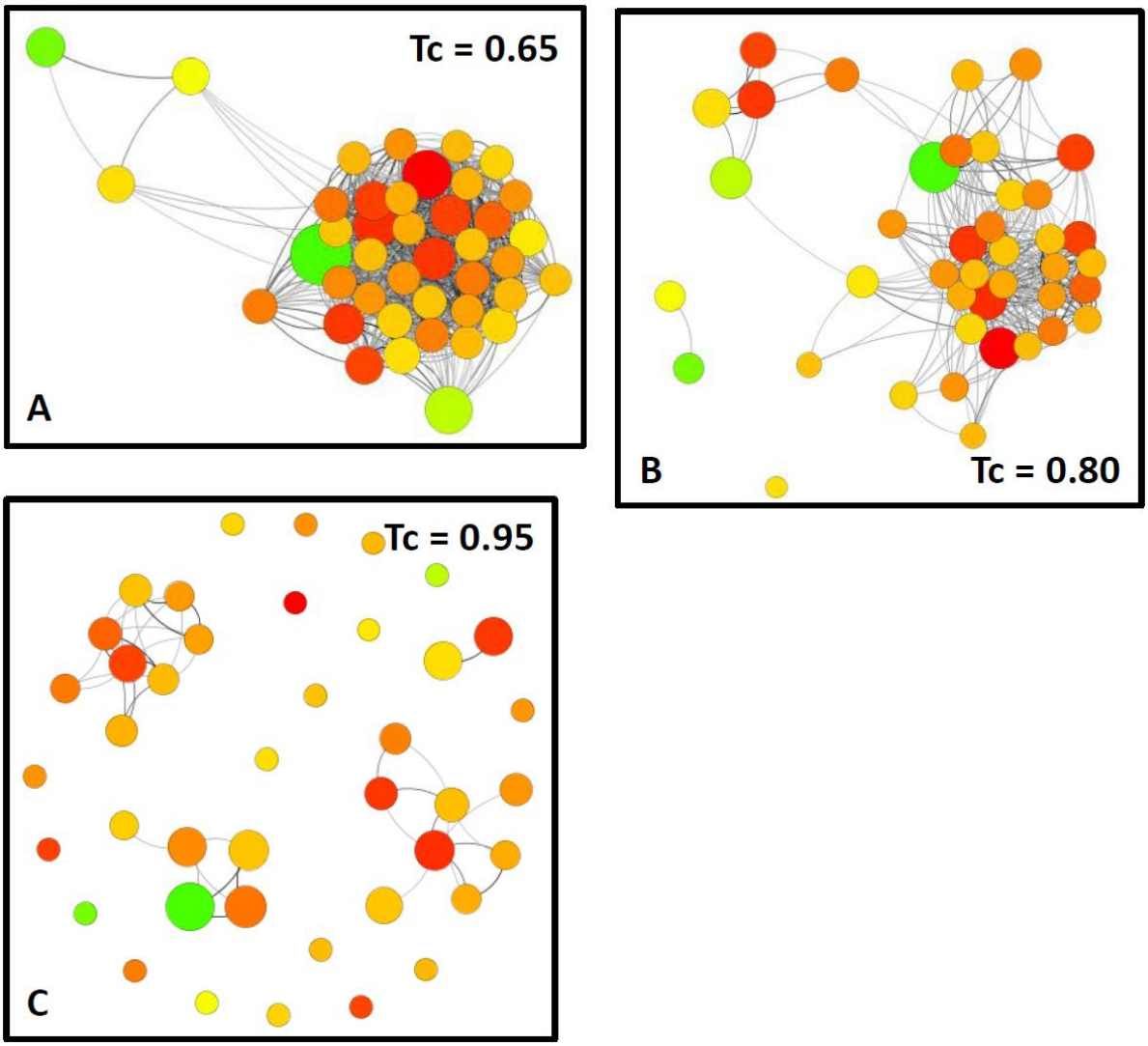


Figure 2: Network-like similarity graphs (NSGs): Topological appearance of the MDR-selectivity data set (KB-3-1/KB-V1) using different MACCS-based similarity thresholds. Similarities are determined using Tanimoto coefficients.

Furthermore, to assign a SAR category to the data set the global SARI was calculated. The SARI score is a numerical function with values in the range from 0 to 1. Low, intermediate and high values reflect three different SAR categories; discontinuous, heterogeneous and

continuous. The global SARI values for the potency networks (KB-3-1 and KB-V1) as well as the selectivity NSG (KB-3-1/KB-V1) were determined (see Table 2).

Table 2: Global SARI scores for potency and selectivity NSGs determined at Tc = 0.65.

	SARI	$score_{cont}$	$score_{disc}$
KB-3-1 NSG	0.547	0.157	0.063
KB-VI NSG	0.539	0.158	0.079
selec NSG	0.602	0.444	0.24

From Table 2 it can be seen that both, the two potency NSGs as well as the selectivity NSG exhibit intermediate SARI scores, which indicates that they follow a heterogeneous SAR. Further investigation of the two subscores revealed that the intermediate SARI scores are based on low $score_{cont}$ and low $score_{disc}$ values. The low $score_{cont}$ indicate the presence of a structural constraint, which is confirmed by the fact that all the thiosemicarbazones under consideration either show modifications at N4 or at N1 and the major structural variation comes from modifications of this two substituents (compare Figure 1). In conclusion the SAR type of this data set can be categorized as: *heterogeneous-constraint SAR*.

Analysis of SSR Pathways in NSGs

The SARANEA software suite also allows the generation of SAR/SSR pathways from NSGs. These pathways organize SAR/SSR information and thereby support the assessment of structural changes that are accompagnied also by changes in activitiy or selectivity. The software systematically computes (see Methods section) pathways from two selected molecules. The graphical representation in the following figures was generated by selecting the compound with the lowest activity/selectivity and the compound with the highest activity/selectivity within each NSG.

Figure 3 displays the shortest pathway within the KB-3-1 NSG from the compound with the lowest activity (Figure 3,1) to the molecule with the highest activity (Figure 3,5). If one keeps in mind that IC_{50} values from the KB-3-1 cell line denote “general cytotoxicity”; i.e. cytotoxicity in absence of the MDR-transporter ABCB1, it can be seen, that thiosemicarbazones without an aromatic substituent at N4 and also a pyridyl substitutent at N1 instead of the β–isatin scaffold (Figure 3; 4,5) are much more potent than thiosemicarbazone analogues with an aryl/benzyl or β–isatin substituent (Figure 3; 1,2,3). Contrary to that, Figure 4 displays the potency NSG of the compounds with their IC50 values derived from the ABCB1-overexpressing KB-V1 cell line.

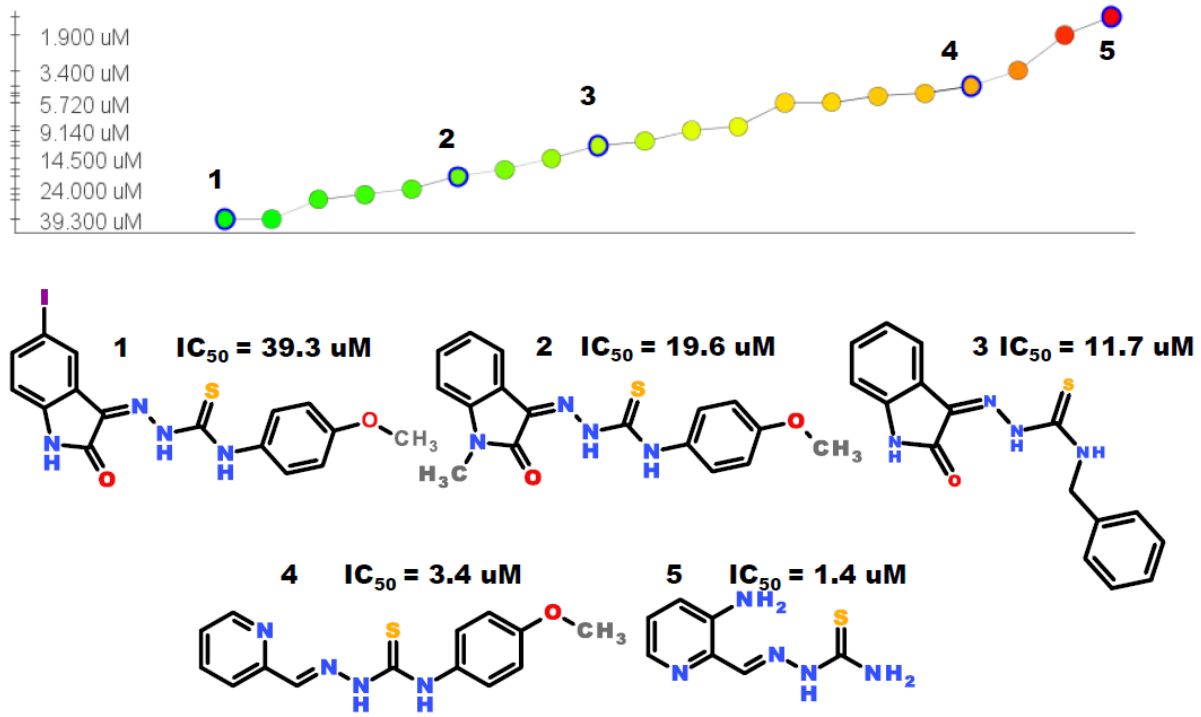


Figure 3: Pathway analysis of the KB-3-1 potency NSG. The color coding denotes the increase in activity (red = highest potency). Color-coded nodes are positioned in the graph according to increasing potency in the KB-3-1 cell line. Molecules corresponding to numbered nodes are depicted as examples in the lower part of the figure.

From the data derived from the ABCB1-overexpressing KB-V1 cell line shown in Figure 4 it can be seen that molecules with both the N4-aryl substitution as well as with the N1-β-isatin scaffold exhibit highest cytotoxic potency in presence of ABCB1. Compound 7 in Figure 4 shows an IC_{50} value of 640.0 nM and constitutes the molecule with the highest potency in this data set. Molecules that contain a substituted β-isatin scaffold and also show an aliphatic modification rather than a cyclic substituent at N4 show IC_{50} values > 10 uM.

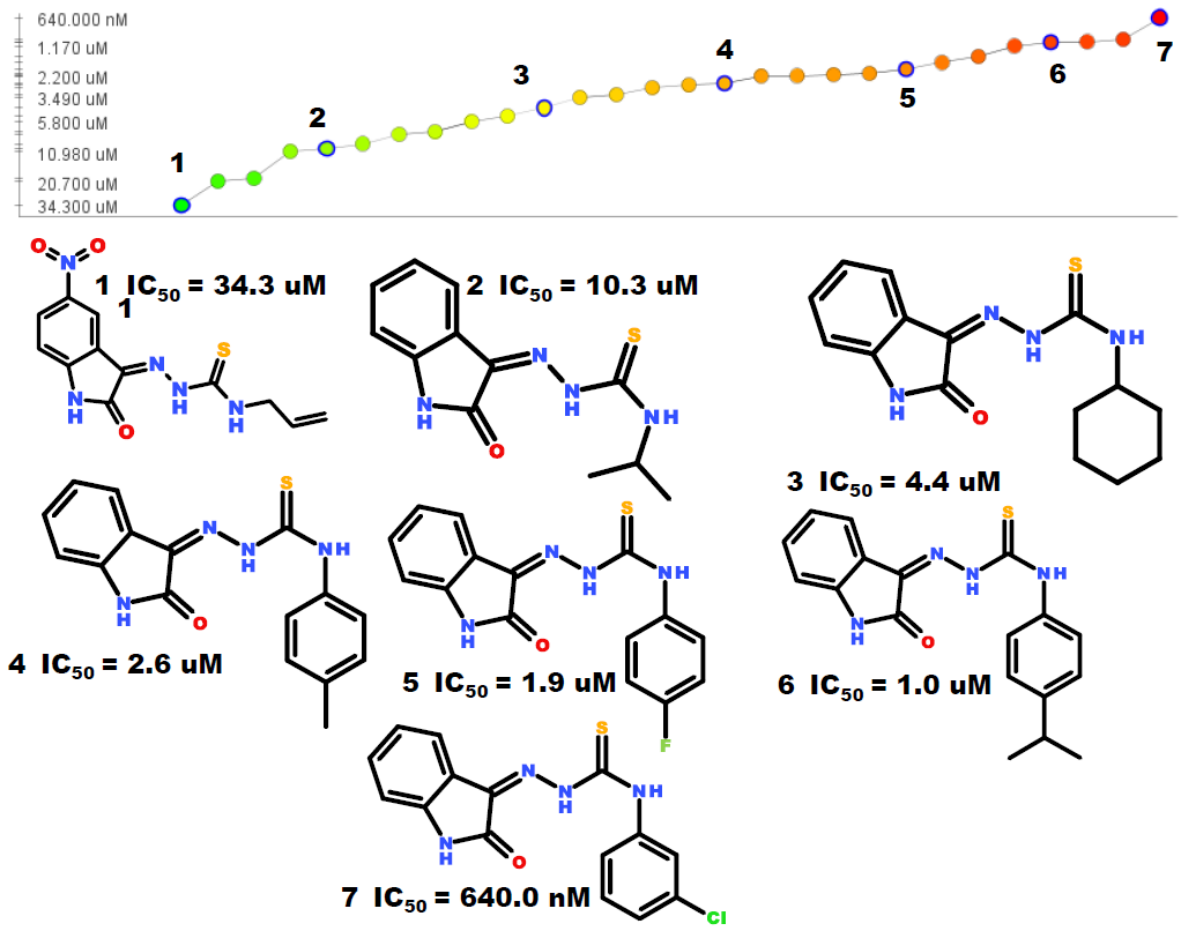


Figure 4: Pathway analysis of the KB-V1 potency NSG. The color coding denotes the increase in activity (red = highest potency). Color-coded nodes are positioned in the graph according to increasing potency in the KB-V1 cell line. Molecules corresponding to numbered nodes are depicted as examples in the lower part of the figure.

The last pathway to be analysed was derived from the selectivity NSG (Figure 5). The shortest pathway calculated from the selectivity NSG that includes both the compound with the smallest SR and also the compound with the highest SR is displayed in the schematic below. Additionally, eight molecules are also displayed as representatives that mirror the gradual structural changes that are accompagnied with changes in selectivity.

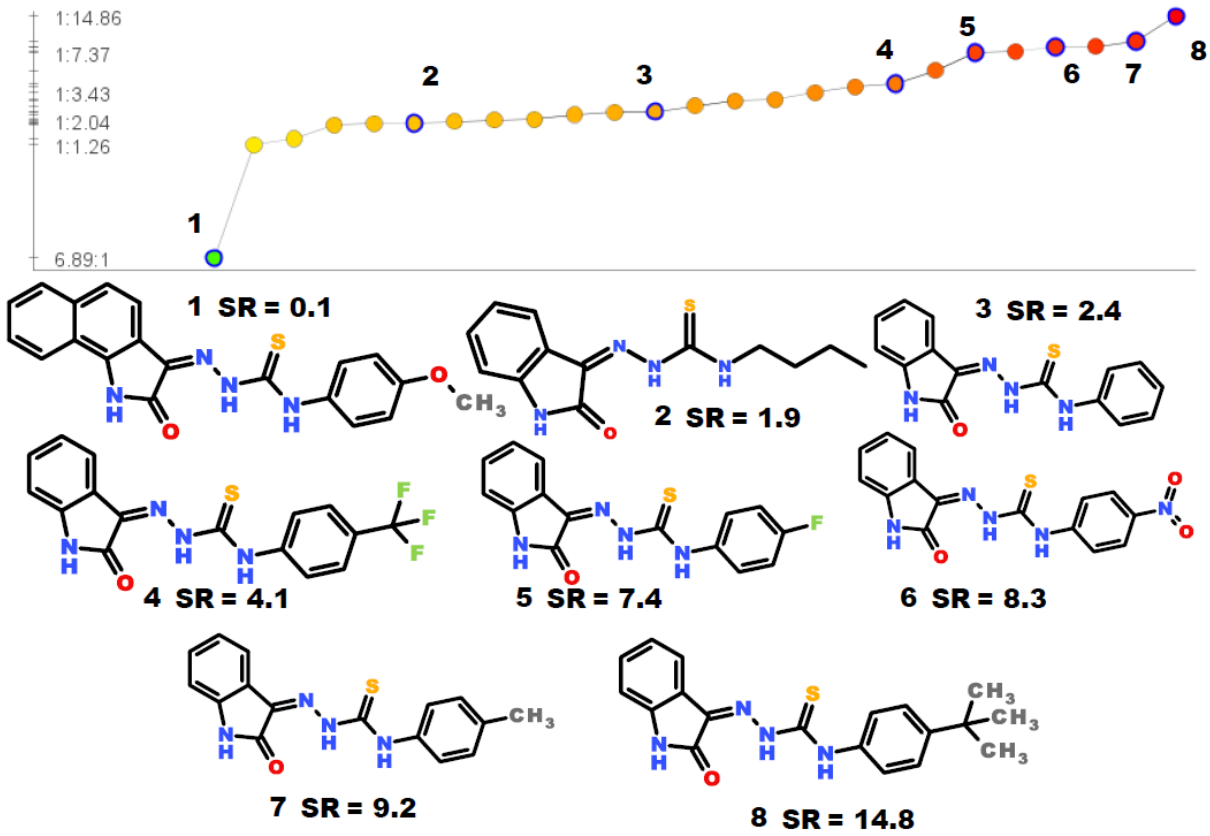


Figure 5: Pathway analysis of the KB-3-1/KB-V1 selectivity NSG. The color coding denotes the increase in selectivity (red = highest SR). Color-coded nodes are positioned in the graph according to increasing selectivity. Molecules corresponding to numbered nodes are depicted as examples in the lower part of the figure.

In general it can be concluded that MDR-selectivity is optimized if molecules contain the β-isatin scaffold as well as a phenyl substituent. Contrary to that, those thiosemicarbazones that are highly active in parental KB-3-1 cells, i.e. exhibit only "general cytotoxicity" show highest potency with an aliphatic substituent at N4 of the thiosemicarbazone and lack the β-isatin substitution. Selectivity towards ABCB1-overexpressing cells is optimized by introducing an aliphatic moiety in para-position of the phenyl ring system. Compound 8 shown in Figure 5 shows a KB-3-1-IC50 =15.9 uM , a KB-V1-IC50 =1.07 uM, which results in a SR of 14.8. This molecule displays the highest potency difference in the data set and is therefore the most MDR-selective from this pharmacological class.

Exploring the chemical neighborhood of the lead compound NSC73306

It is also possible to derive layered chemical neighborhood graphs from NSGs. These graphs serve the utility to explore the environment of a central molecule of interest. According to Türk et al. NSC73306 (1-isatin-4-(4-methoxyphenyl)-3-thiosemicarbazone) is a very promising lead that was shown to selectively kill ABCB1-expressing cells with 4.3-fold selectivity (Türk et al. 2009). In order to explore the chemical neighborhood of the lead compound NSC73306, a radial plot was produced from the selectivity NSG (Figure 6). In this plot molecules (colored nodes) are organized around NSC73306, which is located in the center of Figure 6. Nodes are color-coded (as in the NSG) according to their selectivity and molecules located at concentric circles next to NSC73306 are highly similar to it.

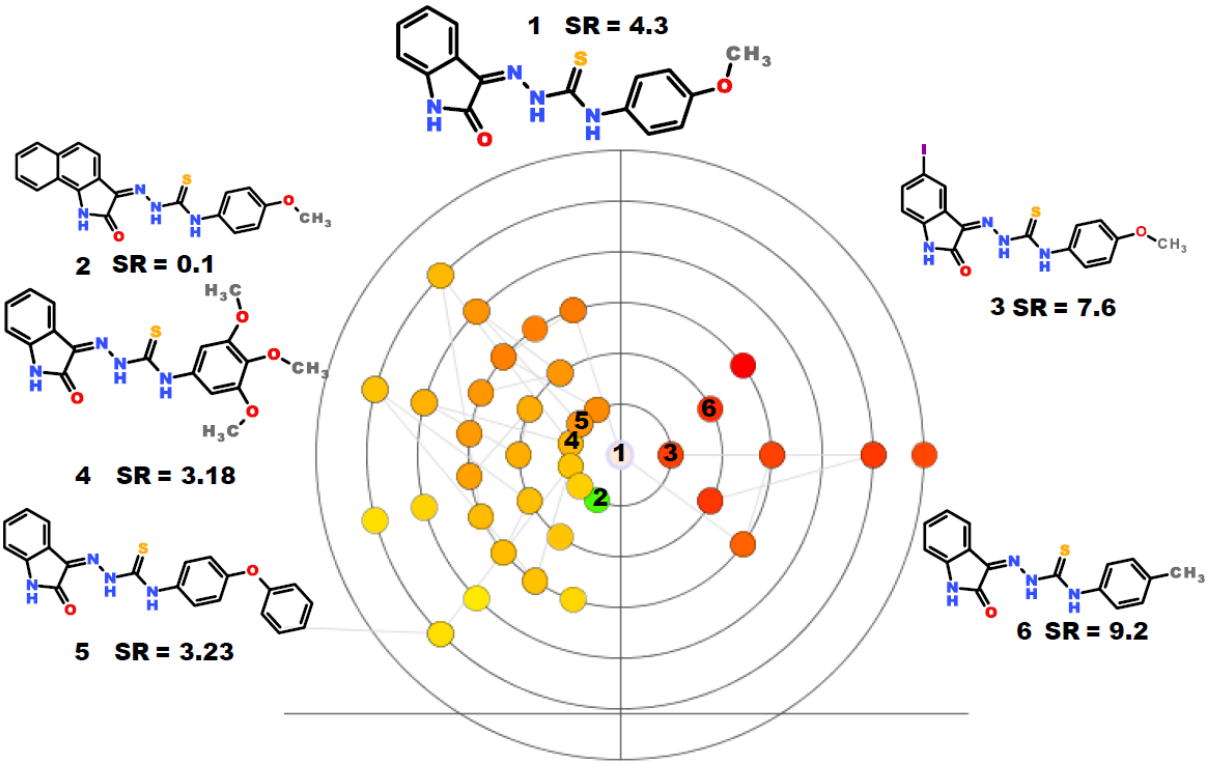


Figure 6: Exploring the chemical neighborhood of the lead compound NSC73306. An layered chemical neighborhood graph is shown for the selectivity NSG. The graph depicts the direct neighbors of NSC73306 (1) at the center of the circle. The concentric circles around the center molecule NSC73306 (1) represent different levels of Tanimoto similarity. Most similar compounds are found at concentric circles next to the center, while more unrelated molecules are positioned at the outermost circle. Molecules located on the left hand side of NSC73306 show less selective molecules (yellow and green nodes), while molecules on the right hand side display higher MDR-selectivity (red nodes).

Figure 6 also displays the chemical structures of direct neighbors of the lead molecule NSC73306. From this radial plot it can be seen that NSC73306 is in close direct proximity to a selectivity cliff, which is constituted by molecule 2 in Figure 6. Molecule 2 in this figure only shows a SR of 0.1, but is chemically highly similar to NSC73306. The only difference is a modification at the β–isatin system. Molecule 2 shows a benzo-annelated indole (β–isatin) system, whereas the β–isatin of NSC73306 does not show any modification. Additionally, molecule 3 which shows a SR of 7.6 is located at the same concentric circle as molecule 2. The fact that NSC73306 is located next to a selectivity cliff marker (molecule 2) suggests, that further chemical optimization in a lead-optimization project must be done very carefully, since small structural changes can easily abrogate the selectivity of this molecule.

Identification and Characterization of a selectivity cliff in the selectivity NSG

At next, this contribution provides a detailed characterization of an observed selectivity cliff in the selectivity NSG. In order to identify selectivity cliff markers, the individual SAR discontinuity score (see Methods section) was computed for all 41 molecules and those molecules with highest discontinuity scores were subjected to further investigation (Figure 7). The analysis of the individual discontinuity scores showed that most molecules show values smaller than 0.5 for the individual discontinuity score (Figure 7A), which is also reflected in the low $score_{disc}$ component of the SARI (see Table 2). Furthermore, we compared the individual discontinuity scores with 2D structural similarity (Figure 7B).

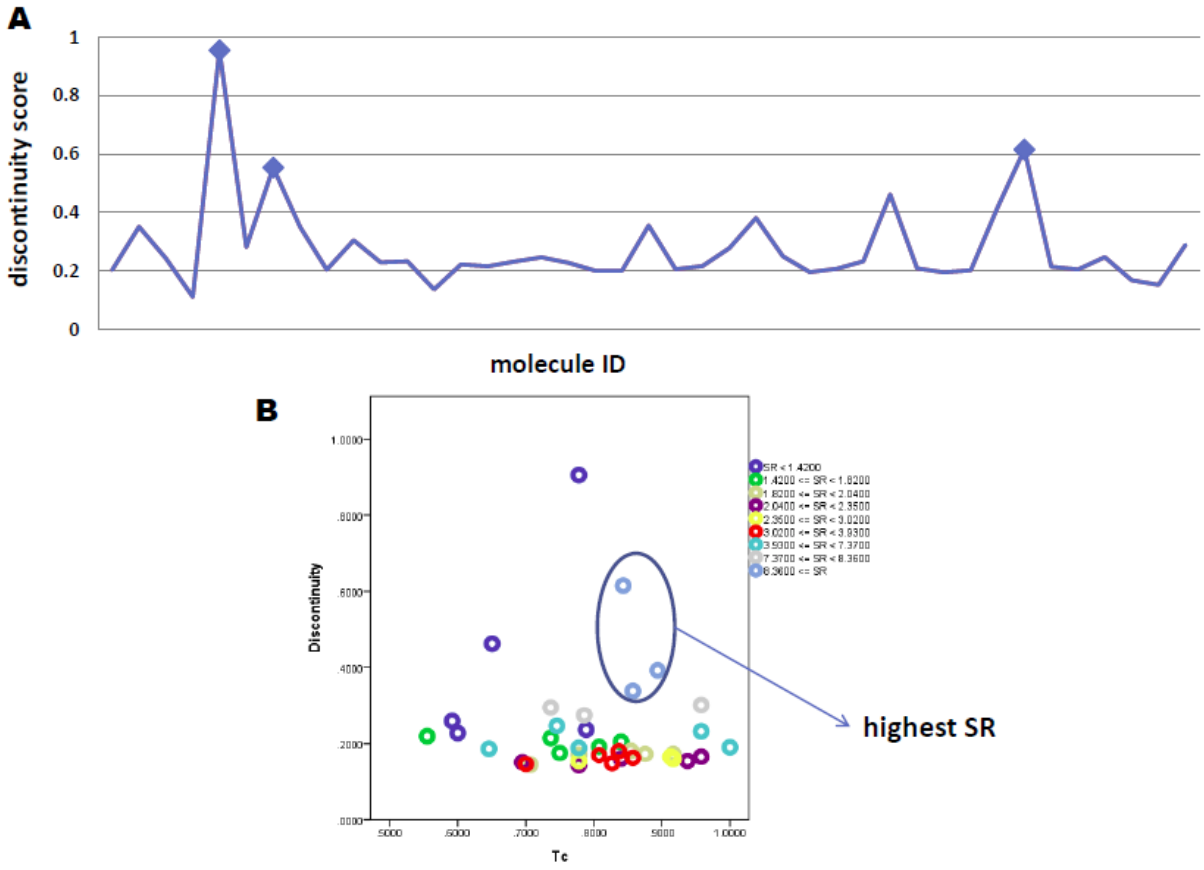


Figure 7: Identification of selectivity cliff markers in the KB-3-1/KB-V1 selectivity NSG using the discontinuity score. A) depicts the calculate individual discontinuity score for the 41 MDR-selective molecules. B) Plot of the individual discontinuity score versus 2D Tanimoto similarity. Data points are color-coded with respect to the observed SR.

From Figure 7B it can be seen that the most selective molecules are among those with the highest individual discontinuity scores. The three molecules with the highest individual discontinuity scores are displayed in Figure 8 alongside with their localization in the selectivity NSG and the two potency NSGs.

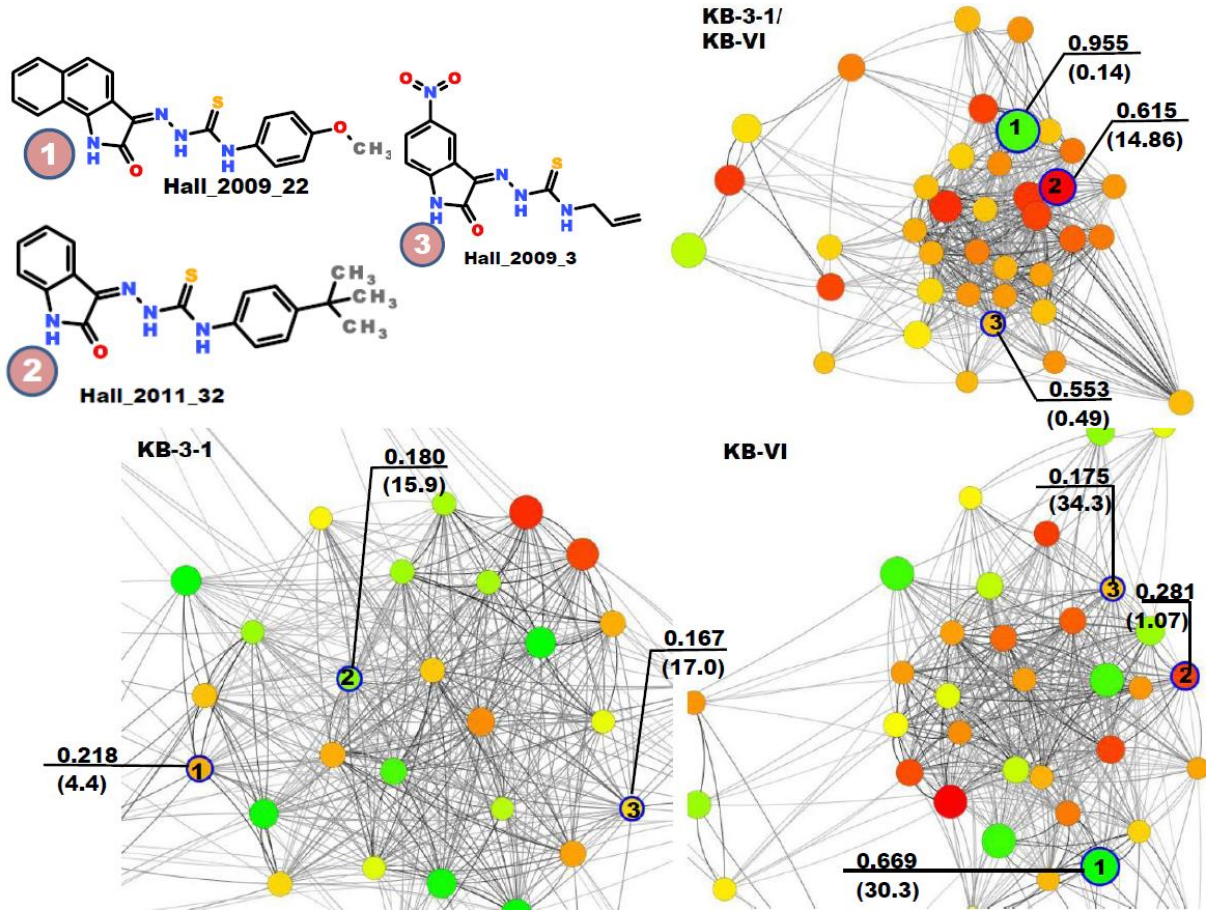


Figure 8: Location of selectivity cliff markers in the selectivity NSG (upper right panel) and the two potency NSGs (lower panel). The three molecules with the highest discontinuity score in the selectivity NSG are also depicted. These three molecules have different potency against and selectivity for the ABCB1-overexpressing cell line. Discontinuity values are shown in bold. Selectivity ratio/IC50 values are shown in brackets. Compound 1 forms an selectivity cliff in the KB-3-1/KB-V1 NSG and also in the KB-V1 potency NSG. The other two molecules exhibit only high discontinuity networks in the selectivity network, while being located in continuous SAR regions in the other networks (small discontinuity values).

Compound 1 in Figure 8 which is also very similar to the lead compound NSC73306 is identified as a selectivity cliff marker.

Identifying Compounds with undesired side effects using toxicity marker sets

In order to get a preliminary idea of the side effect potential of the 41 thiosemicarbazones by means of NSGs, three different toxicity marker sets with known undesired properties were also added to the selectivity NSG at varying Tc thresholds. The organ-focused toxicity sets used here are 64 molecules with known toxicity towards the hematopoietic system, 252 molecules exhibiting known hepatotoxicity and 63 nephrotoxic marker molecules. By adding these marker molecules, the NSG recalculates the molecular similarities for all the molecules under consideration. Those marker molecules that lie within a predefined 2D similarity threshold are connected to the marker molecules via edges and are so topologically displayed in the NSG. The results are graphically depicted in Figure 9. From the figure below it can be seen that only few toxicity markers lie within a Tc > 0.5. In fact, at a Tc > 0.65 only six molecules are connected to any of the 41 MDR-selective molecules.

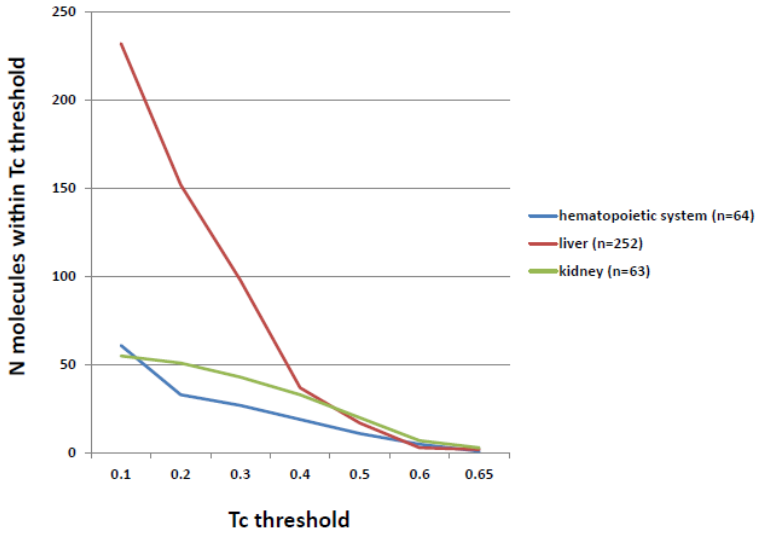


Figure 9: Number of marker molecules connected via edges to nodes of the selectivity NSG in dependency of varying Tc.

The chemical structures of the toxicity markers as well as their topological appearance at Tc > 0.65 are shown in Figure 10. At this threshold only six markers are connected to any of the MDR-selective molecules. Two of the six molecules belong to the liver-toxicity set, one molecule belongs to the hematopoietic system set, and the remaining three toxic molecules belong to the kidney set.

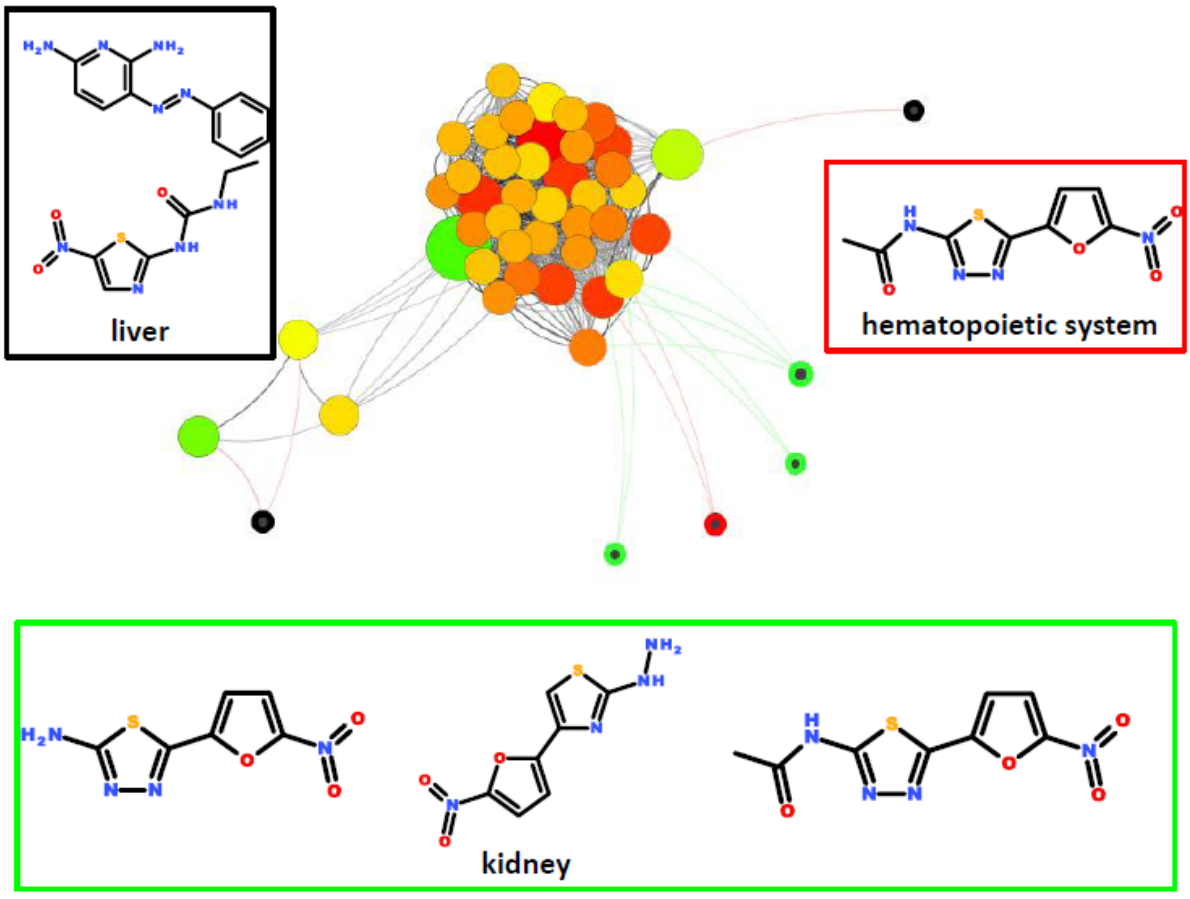


Figure 10: Toxicity marker molecules that show a 2D structural similarity Tc > 0.65 to the 41 MDR selective molecules.

A closer inspection of the chemical nature of these six toxicity markers reveals that five of them contain oxygen hetero atoms and a sulphur atom incorporated into a 5-membered ring system. This five-membered ring system is either constituted of a 1,3-thiazole or a 1,3,4-thiadiazole. The three marker molecules from the liver set as well as the one marker molecule from the hematopoietic system set also contain a 2-nitrofuran system. Only one molecule from the liver-toxicity set does neither contain sulphur nor oxygen. It is only constituted of nitrogen hetero atoms. None of the marker molecules (at Tc > 0.65) is connected to either the lead molecule NSC73306 nor to the most selective molecule with SR = 14.8. In summary, it can be concluded that the series of MDR-selective β-isatin-thiosemicarbazones does not show any significant similarities to a series of known toxico-chemicals. This suggests that these compounds might be safe with respect to toxicity towards the hematopoietic system, the liver, and the kidney. Furthermore, these results go in-line with unpublished data from the Gottesman group in collaboration with Richard Callaghan (narratively described in (Pluchino et al. 2012)). Preliminary animal (dogs and mice) studies with NSC73306 showed that this lead did not exhibit any relevant signs of toxicity at organs known to express endogenous ABCB1 up to doses of 400mg/kg. Nevertheless, these insights must be handled with extreme caution since these results are only based on an *in-silico* analysis of 2D similarities and cannot be regarded to serve as substitute for the usual battery of toxicological in-vitro and animal experiments. However, with respect to future efforts to chemically optimize this compound class, it can be recommended to avoid the introduction of thia-(di)-azoles as well as nitrofurans.

Summary and Conclusion

In this contribution retrospective *in-silico* NSG-modelling was applied to investigate the structural determinants of a series of 41 MDR-selective β–isatin-thiosemicarbazones. MDR-selectivity was defined as the ratio of the IC_{50} values determined in an ABCB1-negative cell line over the IC_{50} values in an ABCB1-expressing cell line. The main results from this investigation can be briefly summarized as follows:

- The primary structural trigger for MDR-selectivity is the β-isatin-thiosemicarbazone scaffold with a phenyl moiety in position N4. The introduction of an aliphatic (ideally t-butyl) *p*-substituent at the phenyl ring increases selectivity up to 14-fold. Lack of either the N4-phenyl ring or the β-isatin system is always associated with higher "general cytotoxicity" (ABCB1-negative KB-3-1 activity) rather than with MDR selectivity.
- The lead molecule NSC73306 was found to be in close proximity to a selectivity cliff. Hence, structural optimization of this molecule must be done with extreme care.
- A preliminary in-silico evaluation of the side effect potential of this novel pharmacological class using 2D structural similarity comparisons, suggests that (on basis of three predefined marker sets) these molecules might eventually harbour a low risk for toxic side effects associated with the hematopoietic system, the liver or the kidney. This further encourages the subsequent exploration of these MDR-selective molecules, but nevertheless must not be confused with a surrogate for pharmacological safety.

If the findings of this contribution can be successfully translated into the further preclinical development of MDR-selective molecules needs to be proven. Future will tell.

Author contribution

I personally collected the chemical and pharmacological information contained in this manuscript. I planned the conceptual design of this contribution and carried out all the calculations. Additionally I am the sole author of this manuscript.

Supplementary Information

ST1:SMILES code, measured activities, SR and individual discontinuity scores for the 41 MDR-selective molecules

SMILES	NAME	IUPAC	KB-3-1 nM	KB-V1 nM	SR	Discontinuity
Cc1cnc(c2c1c(ccc2)N)/C=N/NC(=S)N	Hall_2009_10	[(E)-(5-amino-4-methyl-1-isoquinolyl)methyleneamino]thiourea	1900	2100	0.9	0.23
COc1ccc(cc1)NC(=S)N/N=C\2/c3cc(ccc3NC2=O)I	Hall_2009_18	1-[(Z)-(5-iodo-2-oxo-indolin-3-ylidene)amino]-3-(4-methoxyphenyl)thiourea	39300	5200	7.56	0.27
CN1c2ccccc2/C(=N/NC(=S)Nc3ccc(cc3)OC)/C1=O	Hall_2009_19	1-(4-methoxyphenyl)-3-[(Z)-(1-methyl-2-oxo-indolin-3-ylidene)amino]thiourea	19600	9300	2.11	0.15
COc1ccc(cc1)NC(=S)N/N=C/c2ccccn2	Hall_2009_21	1-(4-methoxyphenyl)-3-[(E)-2-pyridylmethyleneamino]thiourea	3400	2400	1.42	0.22
COc1ccc(cc1)NC(=S)N/N=C\2/c3ccc4ccccc4c3NC2=O	Hall_2009_22	1-(4-methoxyphenyl)-3-[(Z)-(2-oxo-1H-benzo[g]indol-3-ylidene)amino]thiourea	4400	30300	0.15	0.91
COc1ccc(cc1)NC(=S)N/N=C\2/c3cc(ccc3NC2=O)N(=O)=O	Hall_2009_23	1-(4-methoxyphenyl)-3-[(Z)-(5-nitro-2-oxo-indolin-3-ylidene)amino]thiourea	26300	6700	3.93	0.19
C=CCNC(=S)N/N=C\1/c2cc(ccc2NC1=O)N(=O)=O	Hall_2009_3	1-allyl-3-[(Z)-(5-nitro-2-oxo-indolin-3-ylidene)amino]thiourea	17000	34300	0.5	0.46
c1ccc(cc1)NC(=S)N/N=C\2/c3cc(ccc3NC2=O)N(=O)=O	Hall_2009_4	1-[(Z)-(5-nitro-2-oxo-indolin-3-ylidene)amino]-3-phenyl-thiourea	14500	10200	1.42	0.21

c1cc(ccc1NC(=S)N/N=C\2/c3cc(ccc3NC2=O)Br)F	Hall_2009_6	1-[(Z)-(5-bromo-2-oxo-indolin-3-ylidene)amino]-3-(4-fluorophenyl)thiourea	13100	5800	2.26	0.15
c1cc(ccc1NC(=S)N/N=C\2/c3cc(ccc3NC2=O)Br)N(=O)=O	Hall_2009_7	1-[(Z)-(5-bromo-2-oxo-indolin-3-ylidene)amino]-3-(4-nitrophenyl)thiourea	28400	4000	7.1	0.25
c1cc(c(nc1)/C=N/NC(=S)N)N	Hall_2009_9	[(E)-(3-amino-2-pyridyl)methyleneamino]thiourea	1400	5900	0.24	0.26
S=C(Nc1ccc(OC)cc1)[N-]\N=C/1\c2c(NC\1=O)cccc2	Hall_2011_1	1-(4-methoxyphenyl)-3-[(Z)-(2-oxoindolin-3-ylidene)amino]thiourea	14200	3300	4.3	0.19
S=C(NCCCC)[N-]\N=C/1\c2c(NC\1=O)cccc2	Hall_2011_10	1-butyl-3-[(Z)-(2-oxoindolin-3-ylidene)amino]thiourea	10870	5690	1.91	0.15
S=C(NC(C)C)[N-]\N=C/1\c2c(NC\1=O)cccc2	Hall_2011_11	1-isopropyl-3-[(Z)-(2-oxoindolin-3-ylidene)amino]thiourea	39270	10250	3.83	0.18
S=C(NC)[N-]\N=C/1\c2c(NC\1=O)cccc2	Hall_2011_12	1-methyl-3-[(Z)-(2-oxoindolin-3-ylidene)amino]thiourea	35270	19400	1.82	0.18
COc1ccccc1NC(=S)N/N=C\2/c3ccccc3NC2=O	Hall_2011_14	1-(2-methoxyphenyl)-3-[(Z)-(2-oxoindolin-3-ylidene)amino]thiourea	4960	2690	1.84	0.17
S=C(Nc1cc(OC)c(OC)c(OC)c1)[N-]\N=C/1\c2c(NC\1=O)cccc2	Hall_2011_15	1-[(Z)-(2-oxoindolin-3-ylidene)amino]-3-(3-4-5-trimethoxyphenyl)thiourea	11090	3490	3.18	0.15
Clc1ccccc1NC(=S)[N-]\N=C/1\c2c(NC\1=O)cccc2	Hall_2011_16	1-(2-chlorophenyl)-3-[(Z)-(2-oxoindolin-3-ylidene)amino]thiourea	2220	1170	1.9	0.17
Clc1cc(NC(=S)[N-]\N=C/2\c3c(NC\2=O)cccc3)ccc1	Hall_2011_17	1-(3-chlorophenyl)-3-[(Z)-(2-oxoindolin-3-ylidene)amino]thiourea	1870	640	2.92	0.16
Clc1ccc(NC(=S)[N-]\N=C/2\c3c(NC\2=O)cccc3)cc1	Hall_2011_18	1-(4-chlorophenyl)-3-[(Z)-(2-oxoindolin-3-ylidene)amino]thiourea	13360	4990	2.68	0.16
S=C(Nc1ccc(F)cc1)[N-]\N=C/1\c2c(NC\1=O)cccc2	Hall_2011_19	1-(4-fluorophenyl)-3-[(Z)-(2-oxoindolin-3-ylidene)amino]thioure	14150	1920	7.37	0.30

		a				
S=C(Nc1c2c(ccc1)cccc2)[N-]\N=C/1\c2c(NC\1=O)cccc2	Hall_2011_2	1-(1-naphthyl)-3-[(Z)-(2-oxoindolin-3-ylidene)amino]thiourea	5200	2210	2.35	0.17
S=C(Nc1cc(F)c(F)c(F)c1)[N-]\N=C/1\c2c(NC\1=O)cccc2	Hall_2011_20	1-[(Z)-(2-oxoindolin-3-ylidene)amino]-3-(3-4-5-trifluorophenyl)thiourea	4410	2150	2.05	0.17
S=C(Nc1c(F)cc(F)cc1F)[N-]\N=C/1\c2c(NC\1=O)cccc2	Hall_2011_21	1-[(Z)-(2-oxoindolin-3-ylidene)amino]-3-(2-4-6-trifluorophenyl)thiourea	38370	7280	5.27	0.23
S=C(Nc1ccc([N+](=O)[O-])cc1)[N-]\N=C/1\c2c(NC\1=O)cccc2	Hall_2011_22	1-(4-nitrophenyl)-3-[(Z)-(2-oxoindolin-3-ylidene)amino]thiourea	17150	2070	8.29	0.29
S=C(Nc1ccc(O)cc1)[N-]\N=C/1\c2c(NC\1=O)cccc2	Hall_2011_23	1-(4-hydroxyphenyl)-3-[(Z)-(2-oxoindolin-3-ylidene)amino]thiourea	35780	21590	1.66	0.19
S=C(Nc1ccc(N(C)C)cc1)[N-]\N=C/1\c2c(NC\1=O)cccc2	Hall_2011_24	1-(4-dimethylaminophenyl)-3-[(Z)-(2-oxoindolin-3-ylidene)amino]thiourea	6950	2320	3	0.15
S=C(Nc1ccc(Oc2ccccc2)cc1)[N-]\N=C/1\c2c(NC\1=O)cccc2	Hall_2011_26	1-[(Z)-(2-oxoindolin-3-ylidene)amino]-3-(4-phenoxyphenyl)thiourea	5720	1670	3.43	0.17
S=C(Nc1ccc(cc1)C(F)(F)F)[N-]\N=C/1\c2c(NC\1=O)cccc2	Hall_2011_27	1-[(Z)-(2-oxoindolin-3-ylidene)amino]-3-[4-(trifluoromethyl)phenyl]thiourea	8590	2100	4.09	0.19
S=C(Nc1ccc(cc1)C)[N-]\N=C/1\c2c(NC\1=O)cccc2	Hall_2011_28	1-[(Z)-(2-oxoindolin-3-ylidene)amino]-3-(p-tolyl)thiourea	24000	2600	9.23	0.39
S=C(Nc1cc(C)c(cc1)C)[N-]\N=C/1\c2c(NC\1=O)cccc2	Hall_2011_29	1-(3-4-dimethylphenyl)-3-[(Z)-(2-oxoindolin-3-ylidene)amino]thiourea	4370	2200	1.99	0.17
S=C(Nc1cc2CCCc2cc1)[N-]\N=C/1\c2c(NC\1=O)cccc2	Hall_2011_3	1-indan-5-yl-3-[(Z)-(2-oxoindolin-3-ylidene)amino]thiourea	5800	2850	2.04	0.16

S=C(Nc1ccc(cc1)CC)[N-]\N=C/1\c2c(NC\1=O)cccc2	Hall_2011_30	1-(4-ethylphenyl)-3-[(Z)-(2-oxoindolin-3-ylidene)amino]thiourea	4050	1340	3.02	0.16
S=C(Nc1ccc(cc1)C(C)C)[N-]\N=C/1\c2c(NC\1=O)cccc2	Hall_2011_31	1-(4-isopropylphenyl)-3-[(Z)-(2-oxoindolin-3-ylidene)amino]thiourea	9110	1090	8.36	0.34
S=C(Nc1ccc(cc1)C(C)(C)C)[N-]\N=C/1\c2c(NC\1=O)cccc2	Hall_2011_32	1-(4-tert-butylphenyl)-3-[(Z)-(2-oxoindolin-3-ylidene)amino]thiourea	15900	1070	14.9	0.50
S=C(Nc1ccc(cc1)C1CCC(CC1)CC)[N-]\N=C/1\c2c(NC\1=O)cccc2	Hall_2011_33	1-[4-(4-ethylcyclohexyl)phenyl]-3-[(Z)-(2-oxoindolin-3-ylidene)amino]thiourea	17460	10980	1.59	0.18
S=C(Nc1ccccc1)[N-]\N=C/1\c2c(NC\1=O)cccc2	Hall_2011_4	1-[(Z)-(2-oxoindolin-3-ylidene)amino]-3-phenyl-thiourea	3530	1470	2.4	0.17
S=C(NCc1ccccc1)[N-]\N=C/1\c2c(NC\1=O)cccc2	Hall_2011_6	1-benzyl-3-[(Z)-(2-oxoindolin-3-ylidene)amino]thiourea	11680	7630	1.53	0.20
S=C(NC12CC3CC(C1)CC(C2)C3)[N-]\N=C/1\c2c(NC\1=O)cccc2	Hall_2011_7	1-(1-adamantyl)-3-[(Z)-(2-oxoindolin-3-ylidene)amino]thiourea	3160	1020	3.1	0.15
S=C(NC1CCCCC1)[N-]\N=C/1\c2c(NC\1=O)cccc2	Hall_2011_8	1-cyclohexyl-3-[(Z)-(2-oxoindolin-3-ylidene)amino]thiourea	9140	4380	2.09	0.14
S=C(NCC=C)[N-]\N=C/1\c2c(NC\1=O)cccc2	Hall_2011_9	1-allyl-3-[(Z)-(2-oxoindolin-3-ylidene)amino]thiourea	26100	20700	1.26	0.24

References

Djikstra, E.W., 1959. A note on two problems in connexion with graphs. *Numerische Mathematik*, 21, pp.1129–64.

Fruchterman, T.M.. & Reingold, E.M., 1991. Graph drwaing by force-directed placement. *Sotw. Pract. Exper.*, 21, pp.1129–64.

Giacomini, K.M. et al., 2010. Membrane transporters in drug development. *Nature Reviews Drug Discovery*, 9(3), pp.215–36. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20190787 [Accessed March 4, 2012].

Gillet, J.-P. & Gottesman, Michael M, 2012. Overcoming multidrug resistance in cancer: 35 years after the discovery of ABCB1. *Drug Resistance Updates :*, 15(1-2), pp.7762–70. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22465109 [Accessed August 18, 2012].

Goldsborough, A.S. et al., 2011. Collateral sensitivity of multidrug-resistant cells to the orphan drug tiopronin. *Journal of Medicinal Chemistry*, 54(14), pp.4987–97. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3208667&tool=pmcentrez&rendertype=abstract.

Hall, M.D. et al., 2011. Synthesis and Structure Activity Evaluation of Isatin-β- thiosemicarbazones with Improved Selective Activity toward Multidrug-Resistant Cells Expressing P-Glycoprotein. *Journal of Medicnal Chemistry*, 54, pp.5878–5889.

Hall, M.D., Salam, N.K., et al., 2009. Synthesis, Activity, and Pharmacophore Development for Isatin-β-thiosemicarbazones with Selective Activity toward Multidrug-Resistant Cells. *Journal of Medicinal Chemistry*, 52, pp.3191–3204.

Hall, M.D., Handley, M.D. & Gottesman, Michael M, 2009. Is resistance useless? Multidrug resistance and collateral sensitivity. *Trends in Pharmacological Sciences*, 30(10), pp.546–56. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2774243&tool=pmcentrez&rendertype=abstract [Accessed August 18, 2012].

Klepsch, F., Jabeen, I., et al., 2010. Pharmacoinformatic approaches to design natural product type ligands of ABC-transporters. *Curr Pharm Des*, 16(15), pp.1742–52.

Klepsch, F., Stockner, T., et al., 2010. Using structural and mechanistic information to design novel inhibitors/substrates of P-glycoprotein. *Curr Top Med Chem*, 10(17), pp.1769–74.

Longley, D.B. & Johnston, P.G., 2005. Molecular mechanisms of drug resistance. *The Journal of Pathology*, 205(2), pp.275–92. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15641020 [Accessed July 30, 2012].

Lounkine, E. et al., 2010. SARANEA: a freely available program to mine structure-activity and structure-selectivity relationship information in compound data sets. *Journal of Chemical Information and Modeling*, 50(1), pp.68–78. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20053000.

Nobili, S. et al., 2011. OvercomingTumor Multidrug Resistance Using Drugs Able to Evade P-Glycoprotein or to Exploit Its Expression. *Medicinal Research Reviews*, p.Epub ahead of print.

Peltason, L., Weskamp, N., et al., 2009. Exploration of structure-activity relationship determinants in analogue series. *Journal of Medicinal Chemistry*, 52(10), pp.3212–24. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19397320.

Peltason, L. & Bajorath, J., 2007. Quantifying the Nature of Structure−Activity Relationships. *Journal of Medicnal Chemistry*, 50, pp.5571–5578.

Peltason, L., Hu, Y. & Bajorath, J., 2009. From structure-activity to structure-selectivity relationships: quantitative assessment, selectivity cliffs, and key compounds. *ChemMedChem*, 4(11), pp.1864–73. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19750525 [Accessed July 14, 2012].

Pluchino, K.M. et al., 2012. Collateral sensitivity as a strategy against cancer multidrug resistance. *Drug Resistance Updates*, 15(1-2), pp.98–105. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22483810 [Accessed July 25, 2012].

Sarkadi, B. & Szakács, G., 2010. Understanding transport through pharmacological barriers--are we there yet? *Nature reviews. Drug discovery*, 9(11), pp.897–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21031004 [Accessed May 26, 2012].

Shen, D.W. et al., 1986. Multiple drug-resistant human KB carcinoma cells independently selected for high-level resistance to colchicine, adriamycin, or vinblastine show changes in expression of specific proteins. *The Journal of Biological Chemistry*, 261(17), pp.7762–70. Available at: http://www.ncbi.nlm.nih.gov/pubmed/3711108.

Shukla, S., Chen, Z.-S. & Ambudkar, S.V., 2012. Tyrosine kinase inhibitors as modulators of ABC transporter-mediated drug resistance. *Drug resistance updates : reviews and commentaries in antimicrobial and anticancer chemotherapy*, 15(1-2), pp.70–80. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22325423 [Accessed July 25, 2012].

Szakács, G. et al., 2006. Targeting multidrug resistance in cancer. *Nature Reviews Drug Discovery*, 5(3), pp.219–34. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16518375 [Accessed March 10, 2012].

Türk, D. et al., 2009. Identification of compounds selectively killing multidrug-resistant cancer cells. *Cancer Research*, 69(21), pp.8293–301. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2783730&tool=pmcentrez&rendertype=abstract [Accessed August 18, 2012].

Wawer, M. et al., 2008. Structure−Activity Relationship Anatomy by Network-like Similarity Graphs and Local Structure−Activity Relationship Indices. *Journal of Medicinal Chemistry*, 51, pp.6075–6084.

Wawer, M. & Bajorath, J., 2009. Systematic extraction of structure-activity relationship information from biological screening data. *ChemMedChem*, 4(9), pp.1431–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19621333 [Accessed July 14, 2012].

Wawer, M., Peltason, L. & Bajorath, J., 2009. Elucidation of Structure−Activity Relationship Pathways in Biological Screening Data. *Journal of Medicnal Chemistry*, 52, pp.1075–1081.

PART E: CONCLUSION & OUTLOOK

Summary of ligand-based ABC-transporter Machine Learning Models

The focus of this thesis has been to establish predictive machine learning models for ABC-transporter ligands. Major contributions in this thesis were retrieved from the application of ensemble machine learning algorithms. The interpretation of these models helped to discover important properties of compounds interfering with clinical relevant ABC-transporters.

In Chapter 10 a compilation of public available ABCB1-substrates described in the literature (LIT) together with compounds derived from the NCI60 drug sensitivity screen (NCI) was used to explore the capability and feasibility of Friedman`s RuleFit algorithm, an ensemble rule-based method, for modelling the ABCB1 substrate/non-substrate classification problem. The interpretation of this RuleFit model, which showed an overall accuracy of 0.73 in 10-fold crossvalidation, revealed that substrates are larger in size, are characterized by higher hydrophobicity and have more hydrogen-bond acceptors than their respective counterparts. The promising results obtained from RuleFit encouraged us to further explore this classification method. Therefore, we subjected a highly diverse collection of hit and lead molecules (in total more than 1800 molecules) from Boehringer Ingelheim Austria (BIA) to RuleFit modelling (Chapter 12). Here, an additional external training set consisting of more than 1400 compounds was used for assessing the external prediction performance of RuleFit. The molecules of the external validation consisted from the same hit-to-lead projects as the training molecules, but have been synthesized and tested during the progress of this study. For a model based on eleven simple physico-chemical descriptors an excellent prediction performance for the external test set was achieved. However, a similar model based on only a small homogeneous set, failed in predicting the external test set. This undermines the necessity of using a highly diverse library of molecules for modelling ABCB1-substrates. In Summary, RuleFit proved to be a highly useful algorithm for the ABCB1 substrate/non-substrate classification problem that returned models with high predictive performance and was also useful in simplifying the interpretation of the models and supported the identification of molecular properties that characterize potential ABCB1 substrates and non-substrates.

Another ensemble algorithm that was applied in this thesis was Random Forest (RF). RF also retrieved highly predictive models that also provided useful insights into the important features that characterize ABCB1 substrates. In Chapter 13 RF was used to probe the external predictive power of models based on different data sets that were retrieved from literature sources, the NCI screen, and also the BIA library. It was shown that only a diverse collection of BIA training compounds was able to reliably predict the BIA test set. The other models showed an “efficacy-effectiveness” gap; i.e despite good training performance, these models failed to predict a real-life external data set. The application of two different distance-to-model (D2M) measures that aim to assess the applicability domain (AD) of the established models, showed that a distance-based measure outperforms a range-based measure. Furthermore, it was observed that models built on small homogeneous training sets profit more from D2M post-processing than large, diverse (so called global) training sets. However,

the D2M measures were not able to explain why the training set from the literature sources and the training set from the NCI60 database are unable to correctly predict the external test set. A reasonable explanation for this is that the class membership (substrate/non-substrate) of the literature set and the NCI60 set is in both cases based on different pharmacological endpoints than the endpoint of the external test set. Summarizing, "efficacy-effectiveness" gaps can be based on two grounds: differences with respect to the chemical space and differences with respect to the response space.

With respect to data pre-processing two different strategies were pursued in this thesis. One strategy was to explore the effect of feature selection on model performance. Five different FS algorithms were examined with respect to their ability to retrieve useful feature subset for k-nearest neighbor models of ABCB1, ABCC1 and ABCG2 ligands (Chapter 11). It was shown that FS methods which incorporate the classification algorithm gave the best results and returned a small subset of descriptors. Most of the feature sets contained a remarkable quantity of simple counts and van-der-Waals surface area (VSA) descriptors, which is indicative of the general applicability of these descriptor types for models of different ABC-transporter ligands. Certain properties, like molecular size, parital charge and rigidity, have been identified as being useful for discriminating substrates and non-substrates of these three ABC-transporters. For the ABCB1 and the ABCG2 data set cross-validated accuracies of higher than 80% were achieved. Another strategy to prepare the input matrix for modelling tasks is the development of novel descriptors. In this thesis an extension of the SIBAR-approach has been presented (Chapter 14). The so-called MoSS-FP-SIBAR descriptors constitute Tanimoto coefficients calculated from structural fingerprints as feature vectors. These vectors are calculated for a set of reference compounds which contains molecules enriched in class-discriminating structural features. The reference molecules were identified via a Molecular Substructure Search (MoSS). The MoSS-FP-SIBAR descriptors were shown to be useful means to describe a set of recently published ABCB1-substrates and non-substrates. A RF model returned a satisfying performance (MCC: 0.47 in 10-fold CV) on basis of only 15 MoSS-FP-SIBAR descriptors. Furthermore, this model was shown to be directly comparable to a recently published RF model and also to a model based on a 1024-bit fingerprint vector. This showed that the 15 MoSS-FP-SIBAR descriptors introduced herein convey a similar amount of information (with respect to classification performance) than other models based on higher dimensional input data (e.g. the 1024-bit fingerprint model). Furthermore, MoSS-FP-SIBAR descriptors proved to be useful for interpretation purposes. It was shown that the similarities to the two well-known ABCB1 substrates digoxin and domperidone are simple and intuitive means to classify these data.

The last two chapters of this thesis aimed to elucidate the medicinal chemistry of the novel pharmacological class of MDR-selective molecules. The first study made use of 86 molecules that were classified on basis of their selectivity ratio into compounds that selectively kill ABCB1 expressing cells (MDR-selective molecules) and those molecules that were not determined to be selective (non-MDR-selective molecules). In the second study the IC_{50} values were used to denote MDR-selective molecules. In a classification study utilizing 86 MDR-selective and non-MDR-selective molecules it was shown that RF performs best on

a a-priori defined test set (Chapter 15). RF outperformed a gradient boosting machine and a support vector machine. Furthermore, this study featured the feasibility of unsupervised feature selection and returned a feature set of only 30 descriptors out of an initial collection of more than 700 descriptors. The most important descriptors for MDR-selectivity were identified to be autocorrelation vectors weighted by molecular mass. Furthermore, the non-linear behaviour of lipophilicity in MDR-selective molecules has been identified. In Chapter 16 Network-like similarity graphs were employed to identify triggers for MDR-selectivity. Molecules that show a β-isatin substitution as well as a p-phenyl substituted moiety were identified to exhibit highest selectivity ratios. Furthermore, the detailed analysis of the selectivity NSG revealed that the promising lead candidate NSC73306 participates in the formation of a so-called selectivity cliff. Hence, it must be recommended that further chemical optimization of this molecule must be done with caution.

The plethora of models discussed in this thesis, immediately raises the question: *"Which is the best model?"* Definitely, this is not easy to answer. However, when considering internal and external validity, potential for interpretation and also the number of descriptors, it can be concluded that the RuleFit model on the merged BIA data set described by eleven simple physico-chemical descriptors might be a good and reasonable choice.

The following table below briefly summarizes the different contributions to the field contained in this thesis.

Methodological Discussion

In this thesis several different methodological strategies are presented that aim to provide machine learning models of ABC-transporter ligands with good classification performance. The insights from the results achieved herein shall be summarized briefly from a sole methodological viewpoint. In general future efforts can concentrate on several different ways to improve pharmacoinformatic classification models:

(i) improvement of classification algorithms so predictions become more accurate;
(ii) development of new chemical descriptors to further encode chemical information not included in contemporary features commonly used so far
(iii) application of more sophisticated methods to select the most useful set of relevant features from a high dimensional descriptor space
(iv) identifying useful methods that are able to reliably discriminate correct from incorrect classification results, i.e. estimating the applicability domain.

With respect to point (i) it can be summarized on the one hand from the available literature [1–3] but also from the results shown in this thesis (compare to Chapter 12, 13, 14 and 15) that it will be hard to further improve the performance of ensemble algorithms like RuleFit or Random forest (RF). A similar picture occurs when summarizing point (ii) – improvement of chemical descriptors. In Chapter 14 a novel extension of the SIBAR approach is developed, described and critically appraised. The newly introduced MoSS-guided FP-SIBAR descriptors provided useful insights into the chemistry of ABCB1-substrates but fail with respect to substantially improve model performance. Therefore, it can

be concluded that developing novel descriptor types is possible but similar to point (i) improvement over already established algorithms/descriptors will be difficult to achieve in the future. It must be considered that currently the field has achieved some kind of plateau with respect to descriptor or algorithm development.

Qualitative Summary of the Individual Modelling Efforts conducted in this Thesis:

data set	pre-processing/ descriptors	algorithm	post-processing	performance	model information/interpretation	ch.
ABCB1 substrate/non-substrate						
LIT/NCI	11 physico-chemical descriptors	RuleFit	interpretation of ensemble rules	ACC:0.73[§]	*substrate properties:* mr>10 PEOE_VSA_HYD>300 5<logP<7 a_acc>7	10
BIA	11 physico-chemical descriptors	RuleFit	interpretation of ensemble rules	*training:* HOM: 0.67[#] DIV: 0.56[#] MERGED: 0.64[#] *test:* HOM: 0.23[#] DIV: 0.67[#] MERGED: 0.77[#]	*substrate properties:* vsa_acc>37 PEOE_VSA_HYD>400 b_rotN>8 a_don<4	12
NCI	5 FS algorithms	*k*NN	radial visualization plot	ACC:0.85	.) wrappers outperform filter methods *substrate properties:* .)VSA descriptors encoding lipophilicity and size .) high SMR_VSA2; high SlogP_VSA8	11
BIA/LIT /NCI	77 ACORR; 166 MACCS; 32 VSA descriptors	RF	assessment of the applicability domain	*test:* MERGED: 0.79[#] LIT: 0.15[#] NCI: 0.08[#]	.) models on small homogeneous training sets benefit most of D2M measures .) distance-based D2M measures perform better than range-based methods .) data from different resouces are not necessarily directly comparable – different assays reflect different pharmacology	13
Poongavanam et al.[4]	MoSS-FP-SIBAR	RF	interpretation of single decision tree from RF	MCC: 0.47[§]	*most important descriptors:* Tc vectors: digoxin, domperidone *substrate properties:* higher Tc for domperidone and digoxin	14
ABCC1 substrate/non-substrate						
NCI	5 FS algorithms	*k*NN	radial visualization plot	ACC:0.72[§]	.) wrappers outperform filter methods *substrate properties:* .) VSA descriptors encoding partial charges	11
ABCG2 substrate/non-substrate						
NCI	5 FS algorithms	*k*NN	radial visualization plot	ACC: 0.88[§]	.) wrappers outperform filter methods *substrate properties:* .)VSA descriptors encoding lipophilicity and size (SMR_VSA1, SMR_VSA6)	11
MDR-selective molecules						
Türk et al.[5]	UFS	RF GBM SVM	RF variable importance, isoMDS plot for unsupervised RF	*training:* Sens: 0.92 Spec: 0.74 *test:* Sens: 0.71 Spec: 0.92	.) 2D autocorrelation descriptors weighted by molecular mass are useful features to characterize favourable and unfavourable properties of MDR-selective agents	15
Hall et al.[6], [7]	NA	NSG	NA	NA	.) a β-isatin moiety and an additional p-substituted phenyl ring are key for thiosemicarbazone MDR-selectivity .) NSC73306 constitutes a selectivity cliff marker	16

*LIT = public data set compiled from literature sources; **NCI = public data set retrieved from the NCI60 drug sensitivity screen; ***BIA = proprietary data set established and provided by Boehringer Ingelheim Austria; NA = not applicable; ACC= overall accuracy; #results reported as mcc = Matthews correlation coefficient; RF = Random Forest; GBM = gradient boosting machine; § = 10-fold cross-validation; Tc = Tanimoto coefficient; NSG = Network-like similarity graph

Feature selection (FS) – a method that is often applied but sometimes appears to be heavily underestimated with respect to its potential – is another alternative way to improve model performance, but can also be of high usefulness in facilitating model interpretation. However, there is no single, universal FS strategy that performs best on all possible modelling tasks. When comparing the results achieved in Chapter 11 and from the application of the unsupervised feature selection approach presented in Chapter 15 the following conclusion can be drawn. Wrappers, i.e. methods that incorporate the machine learning algorithm directly into the FS process, are definitely best if internal validity is the primary objective and are useful for selecting descriptors that shall be interpreted to understand underlying properties of a data set under consideration. However, it should be considered that such an approach introduces a bias towards the training instances and it must be considered that Wrapper-selected models might perform poorly in external prediction. Contrary, if external prediction is the primary focus, then unsupervised FS strategies, like UFS might be more appropriate (compare Chapter 15). Unsupervised methods weight feature relevance on a total different basis than wrappers. Instead of optimizing prediction performance, they use much simpler measures, like redundancy (e.g. correlation) or information content (e.g. spread (SD)) of the descriptors. In other words, they select features in an unbiased fashion. It must be recommended that descriptor subsets selected by unsupervised methods shall be combined with highly accurate modelling algorithms, e.g. ensemble methods, to achieve good external prediction performance. The field of estimating the applicability domain (point (iv)) comprises definitely the youngest field of these improvement possibilities and also has the potential for adding more progress to improving model performance and reliability. Several useful metrics have been described in the literature so far. In Chapter 14 of this thesis two different methods that characterize the "distance-to-model" of new test instances to the training instances have been explored. However, only limited success was observed for these two metrics, but it can be concluded that distance-based measures might be more appropriate than simple, range-based methods. Furthermore, the results presented in Chapter 14 also demonstrate that an estimation of the applicability domain might be more important if the model is built on a small, homogeneous compound series with limited chemical diversity, rather than a large diverse collection. Similar findings were also observed by Weaver et al. [8]. In Chapter 14 the application of the distance-to-model measure only improved the performance of the homogeneous set and not of the merged or global models.

Furthermore, in more general, it must be considered that "distance-to-model" measures currently suffer from two main disadvantages. First of all, none of the available measures provides information on *when* a predictive model needs to be rebuilt or retrained. Currently the available methods mainly aim to report initial overfitting of the model. Second, many of the measures used so far only make use of chemical information and do not incorporate model information in their assessment of the applicability domain. This can result in serious misleading findings. For instance, a distance-based measure labels a new test set molecule as being out-of-domain on basis of the calculated descriptors. However, the model (for instance a SVM) transforms this descriptor space (e.g. because of its kernel function) in course of the

prediction of the new instance. In such a case the distance-to-model measure becomes obsolete. Therefore, it must be recommended that such measures should not only consider descriptor space but shall also acknowledge model space (in case of a SVM; the transformed feature space). Future efforts on improving methods that try to estimate the applicability domain of predictive models, shall concentrate to overcome these two disadvantages.

At last it is also necessary to comment on the advantages and disadvantages of network-like similarity graphs (NSGs). In Chapter 16 a class of isatin-β-thiosemicarbazones was explored with respect to their selectivity in MDR-overexpressing cells using network-like similarity graphs (NSGs). This exploration outlined important structural determinants that trigger MDR-selectivity and identified a prominent selectivity cliff marker. NSGs were shown to be versatile means to explore structure-activity relationships (SAR) and structure-selectivity relationships (SSR) in a retrospective manner. Hence, the main advantage of NSG is summarized as follows. While other methods are focussing on capturing more general SAR-aspects (e.g. trends between physicochemical descriptors), the application of NSGs presented in this thesis is focused on identifying small and specific structural modifications that have a large impact on activity or selectivity. However, the main limitation of such an approach is that it cannot be applied for prospective purposes, e.g. prediction of activity/selectivity cliff markers. Further methodological efforts that might also incorporate predictive approaches into the analysis of NSGs might additionally advance the application of this method.

Conclusion and Outlook

More than 35 years after the discovery of human ABCB1, the paradigm transporter of the multifaceted protein family of multidrug ABC-transporters, a lot of research has been invested to characterize different types of compounds interfering with this highly polyspecific efflux pump which is detrimental for successful chemotherapy on one hand but also critical for the ADMET profile of novel clinical drug candidates on the other hand. The findings from *in-silico* machine learning approaches contained herein complement previous results in the field and aid in providing a deeper understanding of the molecular basis of ligand-transporter interaction and may also support future drug discovery in reducing attrition rates and increase the speed of innovation. The “holy grail” of future *in-silico* drug discovery is probably to turn the current *data-driven* attitude in the field into a more *prediction-driven* attitude. However, in order to accomplish this aim the field needs to face a lot of different challenges. A very important challenge is from my personal perspective to set clear standards for modelling in order to provide convergence, comparability and above all reproducibility of different *in-silico* modelling efforts. Other areas of pharmaceutical research have already provided clear guidelines that set out the respective standards in their field (e.g.: ICH Q8 – Q10 [9]). It is definitely desirable that guidelines that clearly define quality standards and assure constant quality control of predictive drug discovery models are established in the next future.

References

[1] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, "Random forest: a classification and regression tool for compound classification and QSAR modeling.," *Journal of Chemical Information and Computer Sciences*, vol. 43, no. 6, pp. 786–99, 2003.

[2] V. Svetnik, T. Wang, C. Tong, A. Liaw, R. P. Sheridan, and Q. Song, "Boosting: an ensemble learning tool for compound classification and QSAR modeling.," *Journal of Chemical Information and Modeling*, vol. 45, no. 3, pp. 786–99, 2005.

[3] C. L. Bruce, J. L. Melville, S. D. Pickett, and J. D. Hirst, "Contemporary QSAR classifiers compared.," *Journal of Chemical Information and Modeling*, vol. 47, no. 1, pp. 786–99, 2007.

[4] V. Poongavanam, N. Haider, and G. F. Ecker, "Fingerprint-based in silico models for the prediction of P-glycoprotein substrates and inhibitors.," *Bioorganic & Medicinal Chemistry*, Mar. 2012.

[5] D. Türk, M. D. Hall, B. F. Chu, J. a Ludwig, H. M. Fales, M. M. Gottesman, and G. Szakács, "Identification of compounds selectively killing multidrug-resistant cancer cells.," *Cancer Research*, vol. 69, no. 21, pp. 8293–301, Nov. 2009.

[6] M. D. Hall, N. K. Salam, J. L. Hellawell, H. M. Fales, G. Szakacs, D. E. Hibbs, and M. M. Gottesman, "Synthesis, Activity, and Pharmacophore Development for Isatin-β-thiosemicarbazones with Selective Activity toward Multidrug-Resistant Cells," *Journal of Medicinal Chemistry*, vol. 52, pp. 3191–3204, 2009.

[7] M. D. Hall, K. R. Brimacombe, M. S. Varonka, K. M. Pluchino, J. K. Monda, J. Li, M. J. Walsh, M. B. Boxer, T. H. Warren, H. M. Fales, and M. M. Gottesman, "Synthesis and Structure Activity Evaluation of Isatin-β- thiosemicarbazones with Improved Selective Activity toward Multidrug-Resistant Cells Expressing P-Glycoprotein," *Journal of Medicnal Chemistry*, vol. 54, pp. 5878–5889, 2011.

[8] S. Weaver and M. P. Gleeson, "The importance of the domain of applicability in QSAR modeling.," *Journal of Molecular Graphics & Modelling*, vol. 26, no. 8, pp. 1315–26, Jun. 2008.

[9] "ICH Quality Guidelines," *International Conference of Harmonization*. [Online]. Available: http://www.ich.org/products/guidelines/quality/article/quality-guidelines.html.

APPENDIX I: Poster Conference Contributions

Poster Contribution @ EuroQSAR, 2008, Uppsala, Sweden

COMPARISON OF CONTEMPORARY FEATURE SELECTION ALGORITHMS: APPLICATION TO THE CLASSIFICATION OF SUBSTRATES FOR ABC-TRANSPORTERS

M. Demel[1], O. Krämer[2], P. Ettmayer[2], E. Haaksma[2], A.G.K. Janecek[3], W. Gansterer [3], G.F. Ecker[1]

[1] University of Vienna, Emerging Field Pharmacoinformatics, Department of Medicinal Chemistry, Althanstraße 14, A-1090 Vienna, Austria
[2]Boehringer-Ingelheim Austria GmbH, Medicinal Chemistry/Structural Research, Dr.-Boehringer-Gasse 5-11, A-1121 Vienna, Austria
[3] University of Vienna, Research Lab Computational Technologies and Applications, Lenaugasse 2, A-1080 Vienna, Austria

BACKGROUND

Theory

- In QSAR modelling and classification studies, quantitative representations of molecular structures are encoded in ***information-preserving descriptor (feature) values***
- The selection of the ***most relevant descriptors*** reflecting the ***relationship between chemical structure and biological activity*** is one of the striking challenges in computer-assisted ligand-based drug design
- Feature selection (FS) algorithms can be categorized into two distinct classes: ***filters and wrappers***[1,2]
- Filter methods are fast and classifier-agnostic, whereas wrapper methods use a heuristic to maximize some objective function with respect to a classifier or regression scheme

Medicinal Chemistry

- ***ATP-Binding-Cassette (ABC) transporters*** represent an ubiquitous family of membrane-bound proteins that are mainly responsible for ***conducting chemo-defence mechanisms*** by ***extruding xenobiotics out of living cells***
- The ABC-transporters ***ABCB1*** (P-glycoprotein), ***ABCG2*** (MXR, BCRP), and ***ABCC1*** (MRP1) are notorious for their potential to confer a ***multi-drug resistant phenotype*** to cancer cells
- Furthermore, they are expressed in various tissues and thus ***influence absorption and distribution*** of a broad variety of structurally and functionally unrelated compounds
- Thus, in light of the increasing knowledge on the importance of ABC-transporters for ***bioavailability of candidate compounds***, the focus of interest shifted ***from inhibitor design towards substrate classification***

MATERIAL & METHODS

Data Sets

- Data are derived from the ***NCI60 screen*** on the ABC-transporters ABCB1, ABCC1 and ABCG2[3]

Descriptor Types

- The descriptors used in this study cover all available ***2D descriptors contained in MOE***. Additionally, we utilize ***autocorrelation descriptors, MACCS-key counts, electrotopological state indices***, as well as two different sets of ***SIBAR-descriptors***[4]

Classification Algorithms

- The performance of the applied FS algorithms is evaluated by using a k-nearest-neighbour (kNN) classifier (k=1,3,5,7,9)

Categorization of Applied FS Algorithms

- **Filters that produce feature ranking:**
  - **Information Gain (InfoGain)**[5]: uses the entropy of a given attribute to calculate the InfoGain with respect to the class label and performs feature ranking
  - **RELIEF**[6]: is an instance-based FS algorithm, that performs feature ranking
- **Filter that performs subset selection:**
  - **Correlation-based FS (CFS)** [7]: selects a feature subset, that shows high correlation with the class label,while showing low intercorrelation
- **Wrapper that performs subset selection:**
  - **linSVM**[8]: attributes are selected on basis of the weight assigned by a SVM

AIM OF THE STUDY

In this study we want to investigate the capability of various FS algorithms with respect to their classification accuracy.

RESULTS

Comparison of Filters Producing Feature Ranking: *IG vs. RELIEF (Fig.1)*

- results are shown for the performance of an external test set
- kNN classifier with k=1

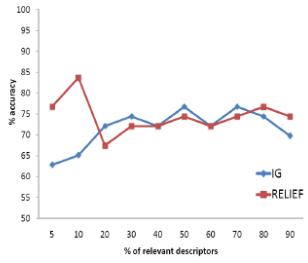


Fig.1A: ABCB1

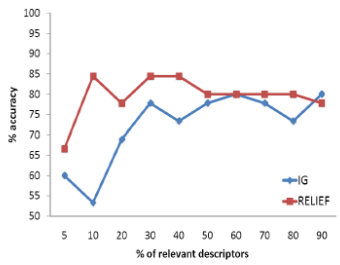


Fig.1B: ABCC1

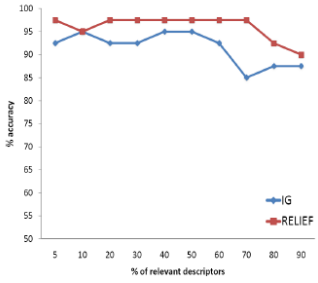


Fig.1C: ABCG2

Comparison of Subset Selection Methods: *CFS vs. WRAPPER (Fig.2)*

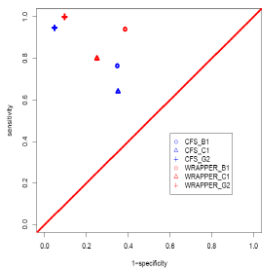


Fig. 2: Model visualization in ROCSpace 2D depiction of the CFS models(blue) and the WRAPPER models(red) on a external test set for kNN (k=1) for details see Ref. 9

Acknowledgements: We gratefully acknowledge financial support from the Austrian Promotion Agency (grant # B1-812074) and from the Austrian Science Fund (grant # L344-N17).

CONCLUSION

Performance of the data sets:

- ✓ **with respect to classification accuracy the following ranking can be achieved for the three ABC-transporters:**
  ABCG2 > ABCB1 > ABCC1

IG vs. RELIEF:

- ✓ **RELIEF outperforms IG; especially when only the top 5% ,10% descriptors are used for model building**

CFS vs. WRAPPER:

- ✓ **The WRAPPER algorithm is superior to the CFS filter method**

References:
1. Demel M., Janecek A., et al., CCADD, 2008.
2. Janecek A., Demel M., et al., JMLR, 2008.
3. Szakacs G., Annereau J.P., et al., Cancer Cell, 2004.
4. Zdrazil B., Kaiser D., et al., QSAR & CombiScience, 2007.
5. Witten I., Frank E., Morgan Kaufman, 2005.
6. Kira K., Rendell L., Ninth International Workshop on Machine Learning, 1992.
7. Hall M., Dept. of Computer Science. Hamilton, New Zealand, 1998.
8. Guyon I., Weston J, et al., Machine Learning, 2002.
9. Fawcett T., Pattern Recogn. Lett., 2006.

Poster Contribution @ EFMC-ISMC, 2008 20th International Symposium of Medicinal Chemistry, Vienna

APPLYING WEB SEARCH ALGORITHMS FOR ADME/TOX PROFILING OF ABCB1 (P-GP) AND HERG LIGANDS

M. Demel[1], A. Janecek[2], R. Schwaha[1], W. Gansterer[2] and G.F. Ecker[1]

[1] University of Vienna, Department of Medicinal Chemistry, Emerging Field Pharmacoinformatics

[2] University of Vienna, Faculty of Computer Science, Research Lab Computational Technologies and Applications

Background

ABCB1 (***MDR1, P-Glycoprotein***), a member of the ABC-transporter drug efflux protein family, as well as the human ether-a-go-go-related-gene (***hERG***) potassium channel are notorious for their potential to invoke ***off-target pharmacologies*** of drugs. Both are membrane-bound proteins, characterised by a ***polyspecific*** ligand recognition pattern. The identification of potential ligands for these two ***anti-targets*** is an important issue for the drug development process in order to reduce ***late-stage failures.***

In this study we adapt ***Latent Semantic Indexing***, a ***similarity-based*** algorithm originally used in ***web search***, ***text mining*** and ***spam filtering***[1] combined with a ***k-Nearest Neighbour*** approach to classify novel drug candidates. We call this approach ***Latent Semantic Drug Classification (LSDC)***.

Methodology

The Basic Vector Space Model (VSM)

- In Information retrieval, documents and queries are often represented as points in ***high-dimensional vector spaces***
- Document vectors are stored in a ***term-document*** matrix ***A***
- ***Distances*** to query vectors are measured in terms of the ***cosines of the angles***
- ***Idea:*** Adapt ***A*** to a ***descriptor-compound matrix***

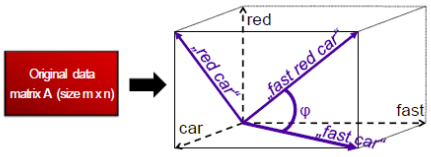


Fig. 1 – Vector space model visualization

Latent Semantic Indexing (LSI)

- Variant of the basic vector space model
- LSI integrates ***automated dimensionality reduction*** based on the ***Singular Value Decomposition*** (SVD, a generalization of ***PCA***)
- LSI works on a ***low rank approximation*** A_k of ***A:*** $A = U\Sigma V^T \approx U_k\Sigma_k V_k^T = A_k$, where k << min(*m*, *n*)
- ***Reduces*** storage requirements *and* computational cost

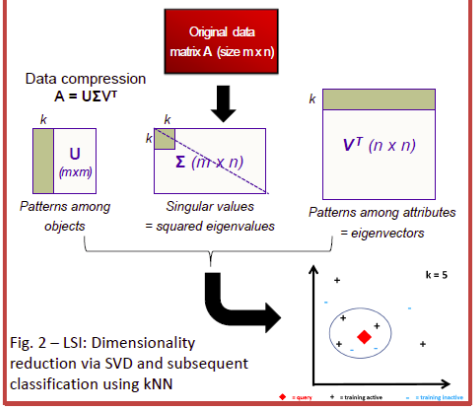
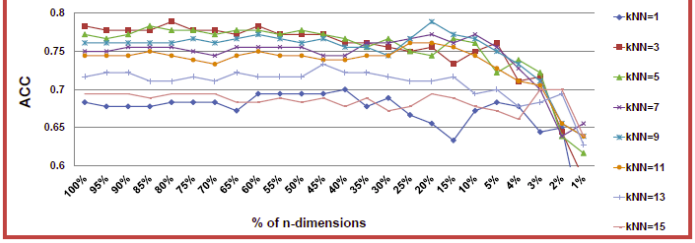


Fig. 2 – LSI: Dimensionality reduction via SVD and subsequent classification using kNN

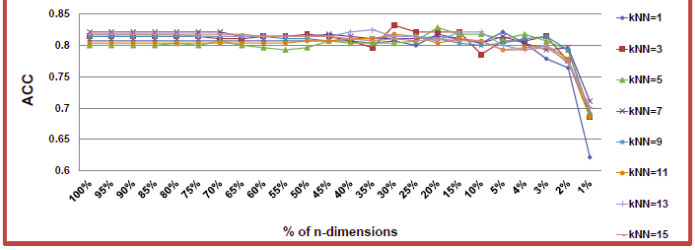
Case Study 1: ABCB1

- LSDC performance on a **249** compound data set collected from the literature [2]
- Compunds are described by **779 MolD2** descriptors
- Results from a 10-fold cross-validation



Case Study 2: hERG

- LSDC performance on a **285** compound data set collected from the literature [3]
- Compunds are described by **779 MolD2** descriptors
- Results from a 10-fold cross-validation



Conclusions:

- **Classification results:** Better for the hERG data set than for ABCB1
- **Dimensionality reduction:** For the hERG data set up to 90%, for the ABCB1 data set up to 70% (for both without significant loss in accuracy)
- ***k*NN:** Smaller numbers of *k* usually show a better performance in both case studies
- **Future work:**

  Explore the capability of LSDC on other antitargets, such as members of the CYP- and nuclear receptor family

  Extend the kNN classification with a similiarity-based threshold

Acknowledgement: We gratefully acknowledge financial support from the Austrian Science Fund (grant # L344-N17), the Austrian Research Promotion Agency (grant # B1-812074), and the CPAMMS-Project (grant # FS397001) in the University Research Area "Computational Science".

References:

1. Gansterer W., A. Janecek, R. Neumayer, *Spam Filtering Based on Latent Semantic Indexing*, in *Survey of Text Mining II*: 2007, Springer
2. Cabrera et al., A topological substructural approach for the prediction of P-glycoprotein substrates, J Pharm Sci., 2006
3. Thai K-M and G.F. Ecker, A binary QSAR model for classification of hERG potassium channel blockers, Bioorg Med Chem, 2008

Poster Contribution @ EFMC-ISMC, 2008 20th International Symposium of Medicinal Chemistry, Vienna

PREDICTIVE MODELS FOR THE CLASSIFICATION OF ABCG2 AND ABCC1 SUBSTRATES/NON-SUBSTRATES TO IMPROVE ADME/TOX PROFILING

Boehringer Ingelheim

M. Demel[1], O. Krämer[2], P. Ettmayer[2], E. Haaksma[2], Janecek A.G.K.[3], Gansterer W.[3], G.F. Ecker[1]

universität wien

[1] University of Vienna, Emerging Field Pharmacoinformatics, Department of Medicinal Chemistry, Althanstraße 14, A-1090 Vienna, Austria
[2] Boehringer-Ingelheim Austria GmbH, Medicinal Chemistry/Structural Research, Dr.-Boehringer-Gasse 5-11, A-1121 Vienna, Austria
[3] University of Vienna, Research Lab Computational Technologies and Applications, Lenaugasse 2, A-1080 Vienna, Austria

Background

❖ Almost 35% of all compounds in the drug development pipeline fail due to ***improper ADMET behaviour***
❖ This fact renders ***in-silico prediction of ADMET properties*** an important issue in drug discovery in order to reduce cost-intense late-stage failure
❖ ATP-binding cassette (ABC) transporters represent an important class of transmembrane proteins responsible for the extrusion of xenobiotics out of mammalian cells
❖ The ABC transporter family members ***ABCG2 (MXR/BCRP)*** and ***ABCC1 (MRP1)*** are notorious for their impact on ***modulating pharmacokinetic properties*** of drugs and play a crucial role in ***chemo-defense*** mechanisms. A prominent example of this is the deployment of ***multi-drug resistance (MDR)*** in cancer cells
❖ Both proteins are characterised by a ***polyspecific*** substrate recognition pattern, i.e. they transport a variety of chemical unrelated, mostly hydrophobic compounds, and ABCC1 additionally extrudes anionic compounds and drug conjugates

Material & Methods

Data sets

❖ Data are derived from the ***NCI60 screen***[1] on the ABC-transporters ABCG2 and ABCC1
❖ This screening data comprise ***Pearson`s correlation coefficients*** of the mRNA levels of the indicated transporter and the cytotoxicity ***as biological activity*** of a test compound
❖ Therefore, we select negative correlated compounds to be substrates and non-correlated compounds to be non-substrates, in order to ***discretize the biological activity*** of our data sets in a ***binary form***

Descriptor type

❖ We calculate ***five different types*** of descriptor sets as input for the classification algorithms used
❖ The descriptor types vary in complexity and interpretability
- **11 ADME descriptors (ADME)**
- **166 MACCS-Atom-type counts (MACCS)**
- **100 SIBAR descriptors (SIBAR)[2]**
- **32 VSA-descriptors (VSA)**
- **77 Autocorrelation descriptors (ACORR)**

Classification algorithms

❖ „single" („base") learners
- **Support Vector Machine (SVM)**
- **k-Nearest Neighbour (kNN)**
- **Decision Tree (DT)**

❖ „Ensemble" learners
- **Random Forest (RF)**
- **Boosting (Boost)**
- **Bagging (Bagg)**

❖ Each of the „base" learners is combinatorially combined with Boosting and Bagging

Characterisation of the data sets

Class distribution and diversity

transporter	N (act)	N (inact)	DI (act)	DI (inact)
ABCG2	94	104	0.27	0.31
ABCG2	103	124	0.32	0.29

Table 1: Distribution of the data sets; N = number; DI=diversity index (on basis of pairwise tanimoto coefficients)

Data sets in PC-(ADME) space

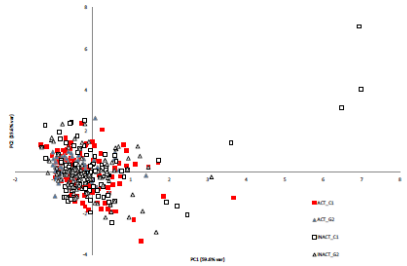


Figure 1: Chemical space of the two data sets; no fill=inact; fill= act; triangles=ABCG2, rectangles=ABCC1

Acknowledgements: We gratefully acknowledge financial support from the Austrian Promotion Agency (grant # B1-812074) and from the Austrian Science Fund (grant # L344-N17).

Classification results ABCG2

❖ external test set performance
❖ results ***for the best models*** are shown in ROC space (for details see ref. 3)

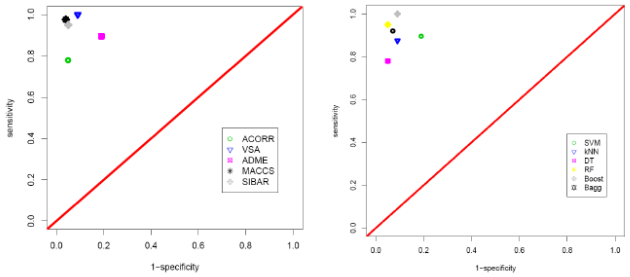


Fig 2: Classification results for the ABCG2 data set; A) regarding the descriptor types, B) regarding the classifiers, red line=random performance

Classification results ABCC1

❖ external test set performance
❖ results ***for the best models*** are shown in ROC space (for details see ref. 3)

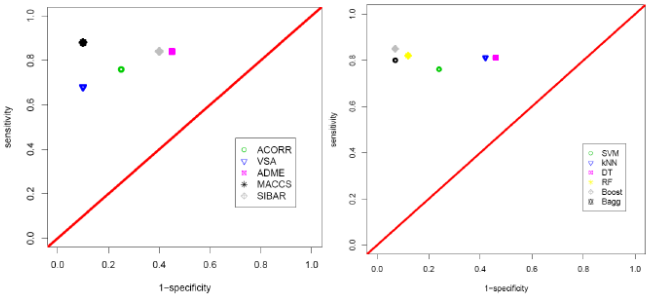


Fig 3: Classification results for the ABCC1 data set; A) regarding the descriptor types, B) regarding the classifiers, red line=random performance

Conclusions

Regarding the data sets:

✓ The ABCG2 data set shows a better performance than the ABCC1 data set

Regarding the descriptor types:

✓ MACCS descriptors show the highest Sensitivity in conjunction with a low false positive rate for both data sets
✓ Contrary ADME descriptors always show a relatively high false positive rate

Regarding the classifiers:

✓ In general ensemble classification algorithms are better than the „base" learners alone (especially Bagg and RF)

References:
1. Szakacs G., et al., Cancer Cell, 2004.
2. Zdrazil B., et al., QSAR & CombiScience, 2007.
3. Fawcett T., Pattern Recogn. Lett., 2006.

Conflict of Interest – Statement

I declare the following conflict of interest:

During the time period 01/11/2006 – 31/03/2009 I was employed as project staff at the Dept. of Medicinal/Pharmaceutical Chemistry of the Faculty of Life Science at the University of Vienna via the FFG grant: #BI-812074.

This grant was financed in part by Boehringer Ingelheim Austria.

Signature Date, Place

Mag. pharm. Michael Demel, MSc – Curriculum Vitae

Date and place of Birth: **23/12/1981 in Vienna, Austria**

Contact: **michael.demel@meduniwien.ac.at**

Work Experience & Education:

01/2013 – current	Qualified Person for Pharmacovigilance *Medical University of Vienna*
01/2013 – current	Lecturer for "Clinical Research" (SE, 2.0 ECTS) FH-Curriculum "Health Assisting Engineering" *FH-Campus: Applied Sciences, Vienna*
08/2012 – current	Lecturer for "Good Clinical Practice" *Medical University of Vienna*
01/2012 – current	GCP-Reviewer for the Ethics Committee of the Medical University of Vienna; *Head: Prof. Ernst Singer* *Medical University of Vienna*
08/2012 – current	Clinical Trial Project Manager *Clinical Trials Coordination Centre* *Head: Prof. Dr. Michael Wolzt* *Medical University of Vienna*
08/2011 – 08/2012	Research Associate at the Medical University of Vienna *Head: Prof. Dr. Michael Wolzt, Dept. of Clinical Pharmacology;* *Medical University of Vienna*
04/2010 – 07/2011	Pharmacist at the "Urania-Apotheke"; *1010 Vienna, Austria* *Stubenring 2*.
01/2011 – 05/2013	Master of Science in Clinical Research; *Graduation with Extinction;* Master Thesis: "Efficacy and Safety of Denosumab in Patients with Advanced Breast Cancer – a randomized, double-blind trial" *Supervisor: PD Dr. Johannes Pleiner-Duxneuner* *Medical University of Vienna, Vienna*
31/03/2010	Accreditation of the "*Austrian Pharmacist`s Diploma"*
04/2009 – 03/2010	Aspirant for the "Austrian Pharmacist`s Diploma" *Opern-Apotheke, 1010 Vienna, Austria*
09/2006 – current	PhD Thesis *Pharmacoinformatics Research Group,* *Head: Prof. Gerhard F. Ecker,* *Dept. of Medicinal/Pharmaceutical Chemistry,* *University of Vienna, Austria*

07/2005 – 08/2006	Master Thesis *Emerging Field Pharmacoinformatics,* *Head: Prof. Gerhard F. Ecker,* *Dept. of Medicinal/Pharmaceutical Chemistry,* *University of Vienna, Austria*
10/2001 – 08/2006	Master of Pharmacy *University of Vienna, Austria*
10/2000 – 05/2001	Compulsory Military Service
09/1992 – 06/2000	High School der Marianisten, Vienna

Signature Date, Place

Abstract (English):

Human ABC-transporters, which act as drug carriers, are notorious for their pivotal role in influencing the pharmacokinetic fate of a plethora of marketed drugs and also for their contribution to MDR, a leading cause of failure of anti-cancer pharmacotherapy in clinical practice. In-silico methods have gained a lot of acceptance in the last years with respect to understand the molecular triggers that drive biological activity of small molecules on the one hand but also with respect to support rational decision making in early phases of drug development on the other hand. In this thesis different machine learning algorithms (Rule-based Modelling, SVM, RandomForests) are employed to characterize proprietary and public data sets of ABC-Transporter substrates and non-substrates. From a pharmacological viewpoint the thesis will concentrate on ABCB1. From a methodological viewpoint the thesis concentrates on the assessment of different feature selection methods, descriptor development (extension of the SIBAR approach), and evaluation of distance-to-model (applicability domain) measurements. An additional focus is also the in-silico characterization of MDR-selective ("collateral sensitive") molecules by means of Network-like Similarity Graphs (NSGs).

Abstract (Deutsch):

Humane ABC-Transporter spielen eine entscheidende Rolle bei der Verteilung von Arzneistoffen im menschlichen Körper und sie sind auch an der Entstehung der multiplen Arzneistoff-Resistenz in Zusammenhang mit Chemotherapie beteiligt. In-silico Klassifikationsmodelle stellen eine schnelle, effiziente und kostengünstige Methode zur identifizierung von ABC-Transporter Substraten dar. In dieser Arbeit kommen verschiedene "Machine Learning Algorithmen" (Rule-based Modelling, RandomForests, Support Vector Machines), die zur Klassifizierung von ABC-Transporter Substraten und Nicht-Substraten dienen, zur Anwendung. Von einem methodischem Standpunkt werden unterschiedliche "Feature Selection"-Methoden, neue chemische Deskriptoren und Maße zur Abschätzung der "Applicability Domain" evaluiert. Zusätzlich wird noch die neue pharmakologische Klasse der sog. "MDR-selective ("collateral sensitive") compounds" mithilfe von Network-like Similarity Graphs (NSGs) charakterisiert.