Machine Learning for Modelling Tissue Distribution of Drugs and the Impact of Transporters

Natália Luísa de Moura Aniceto

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF REQUIREMENTS
OF THE UNIVERSITY OF KENT AND THE UNIVERSITY OF GREENWICH
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

September 2017

DECLARATION

I certify that this work has not been accepted in substance for any degree, and is not concurrently being submitted for any degree other than that of Doctor of Philosophy being studied at the Universities of Kent and Greenwich. I also declare that this work is the result of my own investigations except where otherwise identified by references and that I have not plagiarised the work of others.

PhD Candidate:	Date:
Natáliaaniceto	_20_ / _04_ / _2018_
Natália Aniceto	

ACKNOWLEDGEMENTS

First I would like to thank my supervisors Dr. Taravat Ghafourian, Prof. Alex Freitas and Dr. Andreas Bender for inspiring me, through their own example, to always do my best. Thank you for your help and support, which allowed me to enjoy working throughout my PhD, and thank you for providing me with the amazing opportunity to work with, and learn from you. Additionally, I thank Dr. Vadim Sumbayev for taking care of the logistical side of my PhD during the final year.

To my grandfather, who passed away the year I started my PhD – he will always remain a humbling reference of the real pursuit of knowledge which depends uniquely on the drive to know and not on diplomas and accolades. Despite having just received the basic level of education he was one of the brightest and most inquisitive minds I have ever met.

To my parents, thank you for never stopping pretending you are interested when I explain my research, and for always trying to read my papers. I owe my creativity to my dad and I owe my pragmatism to my mom. I have found that, in science, one cannot exist without the other and, for that, I will always be thankful to you.

To my brother – thank you for always challenging my mind with discussions about almost everything. You always remind me of the pleasure that comes from understanding the phenomena that surround us, and I truly admire you for that.

To Nuno, my partner in all adventures, who is an exquisite scientist - I admire you for your relentlessness in the face of the difficulty. You have inspired me by your ability to persist even when the task is daunting. You have a creativity beyond any boundaries and a mind that keeps me on my feet. Thank you for encouraging me to opt for the “research project” option during our 4th year in university – without this I would quite possibly never discover my love for science.

To my closest friends – no need to name you – some of you I have known for as long as I know myself and some of you are more recent acquisitions, but all of you are important in irreplaceable ways. Thank you for keeping me grounded and sane.

To all the giants upon whose shoulders I stand – thank you.

ABSTRACT

The ability to predict human pharmacokinetics in early stages of drug development is of paramount importance to prevent late stage attrition as well as in managing toxicity. This thesis explores the machine learning modelling of one of the main pharmacokinetics parameters that determines the therapeutic success of a drug – volume of distribution. In order to do so, a variety of physiological phenomena with known mechanisms of impact on drug distribution were considered as input features during the modelling of volume of distribution namely, Solute Carriers-mediated uptake and ATP-binding Cassette-mediated efflux, drug-induced phospholipidosis and plasma protein binding. These were paired with molecular descriptors to provide both chemical and biological information to the building of the predictive models.

Since biological data used as input is limited, prior to modelling volume of distribution, the various types of physiological descriptors were also modelled. Here, a focus was placed on harnessing the information contained in correlations within the two transporter families, which was done by using multi-label classification. The application of such approach to transporter data is very recent and its use to model Solute Carriers data, for example, is reported here for the first time. On both transporter families, there was evidence that accounting for correlations between transporters offers useful information that is not portrayed by molecular descriptors. This effort also allowed uncovering new potential links between members of the Solute Carriers family, which are not obvious from a purely physiological standpoint.

The models created for the different physiological parameters were then used to predict these parameters and fill in the gaps in the available experimental data, and the resulting merging of experimental and predicted data was used to model volume of distribution. This exercise improved the accuracy of volume of distribution models, and the generated models incorporated a wide variety of the different physiological descriptors supplied along with molecular features. The use of most of these physiological descriptors in the modelling of distribution is unprecedented, which is one of the main novelty points of this thesis.

Additionally, as a parallel complementary work, a new method to characterize the predictive reliability of machine learning classification model was proposed, and an in depth analysis of mispredictions, their trends and causes was carried out, using one of the transporter models as example. This is an important complement to the main body of work in this thesis, as predictive performance is necessarily tied to prediction reliability.

CONTENTS

LIST OF PUBLICATIONS

The full list of peer-reviewed publications and scientific presentation that resulted from the research carried out in this thesis is detailed below.

Peer-reviewed Articles

Aniceto N, Freitas AA, Bender A, Ghafourian T. Simultaneous prediction of four ATP-binding cassette transporters substrates using multi-label QSAR. Molecular Informatics. **2016**;35(10):514–28.

Aniceto N, Freitas AA, Bender A, Ghafourian T. A novel applicability domain technique for mapping predictive reliability across the chemical space of a QSAR: reliability-density neighbourhood. Journal of Cheminformatics. **2016**; 8(1)**,** 69.

Articles Under Preparation for Peer-Reviewed Publication

Aniceto N, Freitas AA, Bender A, Ghafourian T. Uncovering new relationships between members of the Solute Carrier superfamily. Journal of Chemical Information and Modelling. Manuscript under preparation.

Aniceto N, Freitas AA, Bender A, Ghafourian T. Impact of Membrane Transporters and Phospholipidosis in Pharmacokinetics: Prediction of Human Volume of Distribution. Journal of Chemical Information and Modelling. Manuscript under preparation.

Aniceto N, Freitas AA, Bender A, Ghafourian T. Modelling of Volume of Distribution by accounting for transporter binding, transporter tissue expression, phospholipidosis and plasma protein binding. Journal of Chemical Information and Modelling. Manuscript under preparation.

Poster Presentations

Aniceto N, Freitas AA, Bender A, Ghafourian T. Accounting for transporter interaction in the prediction of ATP binding cassette efflux using multi-label classification models. **2015**. UKQSAR Autumn meeting, Cresset. **(Best poster prize)**.

Aniceto N, Freitas AA, Bender A, Ghafourian T. Prediction of various ATP-binding cassette substrates via decision trees and multi-label classification. **2015.** 11th German Conference on Cheminformatics.

Aniceto N, Freitas AA, Bender A, Ghafourian T. Reliability-Density Neighbourhood (RDN): addressing the locality of training space to improve Applicability Domain characterization. **2016**. 7th Joint Sheffield Conference on Chemoinformatics

Aniceto N, Freitas AA, Bender A, Ghafourian T. Discovering predictive relationships between members of the Solute Carrier superfamily with multi-label classification. **2016.** 12th German Conference on Cheminformatics.

Oral Presentations

Improved Applicability Domain Determination of QSAR Models Using Local Mapping: RDN. **March 2016**. GlaxoSmithKline, Stevenage, UK.

Novel methods for confidence analysis of predictive models. **January 2017**. Lhasa Limited, Leeds, UK.

LIST OF FIGURES

LIST OF TABLES

Table 6.2. Predictive accuracy on the validation set, using ePL. The number of compounds in the training and validatiom sets were 398 and 133 respectively. The two best models (selected for further analysis) from both Table 6.2 and 6.3 are highlighted in bold. Regarding the feature content available in pre-processing, “all feature” corresponds to 315 features being made available, “MDs” corresponds to 304 features, “PDs” corresponds to 11 features, and “FS-MDs + FS-PDs” corresponds to separate feature selection procedures performed on 304 MDs and 11 PDs. FS: feature selection. *both models resulted from the same input feature set, hence same performance.

Table 6.3. Predictive accuracy on the validation set, using pPL. The two best models (considering both Table 6.2 and 6.3) are highlighted in boldface. The selection of these two models was based on the lowest MAE among all available models. The number of compounds in the training and validation sets were 398 and 133 respectively. *both models resulted from the same input feature set, hence same performance.

Table 6.4. Predictive performance of the two best models on the test set.

Table 6.5. Frequency of combination sizes of physiological descriptors occurring in the same rule. The rules where these combinations occur may or may not contain molecular descriptors as well. 64% of the full collection of if-then rules contain at least one PD.

Table 6.6. Summary of the results obtained by the M5 model tree built with different sets of descriptors. The best performance values are highlighted in bold and underlined. This exercise tests the ability of PDs being selected in a harsher embedded feature selection environment, and is not meant to create alternative (competitive) models to the RF and BT models.

Table 6.7. Summary of predictive performance from Gombar and Hall (Gombar and Hall, 2013) and this work (model 8a), evaluated on an external dataset (N = 30).

Table 6.8. Summary of predictive performance measures from Lombardo and Jing (Lombardo and Jing, 2016) and this work, evaluated on an external dataset (N = 34).

Table 6.9. Summary of predictive performances from the different variants of the Vss modelling conditions. All performances result from testing the models on a fixed, common dataset.

Table 7.1. Summary of the internal validation performance of the various modelling conditions tested. They are compared with the best model obtained in Chapter 6 (named there as model 8a), here identified as “best previous” in this current Table. All models under the same modelling block as the best model (i.e. regression + feature selection conditions that produced the best model) are here considered as the baseline, and identified as such in this Table. Here, “baseline” means the best scenario using no form of transporter expression correction. The two best models, selected based on the lowest MAE are highlighted in boldface. The downward arrows indicate an improvement against the equivalent baseline (no correction) models.

Table 7.2. Summary of predictive performance measured on the test set for the best model in this work (M5) and the best model from Chapter 6 (8a).

Table 7.3. Full list of descriptors used in the best model (M5), and their relative importance (i.e. percentage of correctly predicted training compounds, over the total number of training compounds).

Table 7.4. Comparison of predictive performance between the current best model (M5), the models by Gombar and Hall (Gombar and Hall, 2013) and the previous best model (8a), evaluated on a common external benchmark dataset (N = 30). SVM and MLR stand for support vector machine and multiple linear regression respectively.

LIST OF SCHEMES

ABBREVIATIONS

ABC: ATP-Binding Cassette

Acc: Accuracy

AC(s): activity cliff(s)

AD: applicability domain

ADME: Absorption, Distribution, Metabolism and Excretion

AM1: Austin model 1

ATP: Adenosine Triphosphate

bACC: balanced acuracy

BBB: Blood-Brain Barrier

BCRP (or BCRP1): Breast Cancer Resistance Protein

BCT: Boosted classification trees

BR: binary relevance

BT: boosted trees

CAS: Chemical Abstracts Service (identifier)

CC: classifier chain

CFS: correlation-based feature selection

CID: PubChem Compound Identifier

CV: cross validation

dk-NN: density k-nearest neighbours

DTM: distance-to-model

e{feature} : experimentally derived feature (where the name of the feature replaces "{feature}")

EMA: European Medicines Agency

ER: efflux ratio

ECFP: extended connectivity fingerprints

Fb: Bound fraction

FDA: Food and Drug Administration

FE: fold error

FN: false negative

FP: false positive

FS-All: feature selection (applied to) All (features)

FS-MDs: feature selection (applied to) molecular descriptors

FS-PDs: feature selection (applied to) hysiological descriptors

Fu: Unbound fraction

Fup: unbound fraction of drug in plasma

Fut: unbound fraction of drug in tissue

GA: genetic algorithm (search)

G-mean: geometric mean between Sen and Spe

GMFE: Geometric Mean Fold Error

GS: greedy (stepwise) search

HL: Hamming Loss

IV: internal validation (set)

KDE: kernel density estimation

kNN: k nearest neighbours

LP: label powerset

MAE: mean absolute error

MCC: Matthew's correlation coefficient

MDs: molecular descriptors

MDS: Multidimensional Scaling

MFE: Mean Fold Error

MNDO: modified neglect of differential overlap

MOE: Molecular Operating Environment

MRP1: Multidrug Resistance Protein 1

MRP2: Multidrug Resistance Protein 2

MW: Molecular Weight

NS: non-substrate(s)

OATP: Organic Anion-Transporting Polypeptide

OCT1: Organic Cation Transporter 1

p{feature}: prediction-derived feature (where the name of the feature replaces "{feature}")

PD: Pharmacodynamics

PDs: physiological descriptors

PEOE: Partial Equalization of Orbital Electronegativities

PEPT1: Peptide Transporter 1

P-gp: P-Glycoprotein

PK: Pharmacokinetics

PM3: Parameterization Method 3

PM6: Parameterization Method 6

PPB: plasma protein binding

p{feature}_c : prediction-derived feature, corrected (where the name of the feature replaces “{feature}”

QSAR: quantitative structure-activity relationship

R^2: coefficient of determination

RDN: reliability density neighbourhood

REACH: European Chemical Regulation

Regarding the boosted regression trees (BRT)

RF: random forests

RMSE: Root Mean Squared Error

S: substrate(s)

Sen: sensitivity

SLC: Solute Carrier

SMILES: simplified molecular-input line-entry system

SMOTE: Synthetic Minority Over-sampling Technique

SMP: slope mismatch penalty

Spe: specificity

STD: standard deviation

SVM: support vector machines

{t}-BR: transporter t under the Binary Relevance model

{t}-CC: transporter t under the Binary Relevance model

t-SNE: t-Distributed Stochastic Neighbor Embedding

TE: test (set)

TMD: transmembrane domain

TN: true negatives

TP: true positives

TR: training (set)

Vd: volume of distribution

Vp: plasma volume

Vss: volume of distribution at stationary state

Vt: tissue volume

ΔPR: difference between precision and recall

1. Introduction - Part I: Volume of Distribution and Drug Discovery

1.1. Relevance of ADME to Efficiency within the Pharmaceutical Industry.

The development of combinatorial synthesis capabilities that started in the 1980s was responsible for an estimated 800-fold increase in the amount of compounds synthesized in a year (Scannell et al., 2012). This allowed the generation of hundreds of thousands of compounds in a short period of time and, as a result, was expected to boost efficiency in drug discovery. However, this expectation was not observed, as in the last two decades the number of new compounds brought to market did not increase, with one of the main reasons being undesirable Absorption, Distribution, Metabolism and Excretion (ADME) characteristics (Tian et al., 2015, Cook et al., 2014).

In fact, as of 1997, the reported portion of clinical failures due to pharmacokinetics alone was 39% (Kennedy, 1997, Prentis et al., 1988), being reported in several works as the main cause for failure around this time period (van de Waterbeemd and Gifford, 2003, Kennedy, 1997, Waring et al., 2015, Prentis et al., 1988). However, a recent analysis of AstraZeneca's pipeline between 2005 and 2010 shows pharmacokinetics/pharmacodynamics (PK/PD) issues have decreased immensely, with only 15% of failures being attributed to PK/PD issues in phase I (Cook et al., 2014). In line with this, another recent analysis of four of the main pharmaceutical companies (Waring et al., 2015), referring to the same period of time, shows a 25% failure rate ascribed to PK/bioavailability in phase I. Similar to what was observed for the AstraZeneca's cohort, here too the period of 2006 to 2010 shows that only 10% of analysed compounds (N=243) have failed due to these types of issues. Such improvement can be partly attributed to the fact that, in a way to address and reduce such large attrition rates associated to PK issues, ADME properties started to be evaluated earlier in the drug discovery process, which prompted the need for large-scale screening methods (Boyer et al., 2015). One of the steps towards scaling up ADME screening was the implementation of plate-based in vitro assays to measure key PK properties. This was followed by the integration of in silico tests to the already used in vitro and in vivo tests (Honorio et al., 2013, Kharkar, 2010, Peakman et al., 2015). Indeed, looking at AstraZeneca's 2005-2010 cohort data, prior utilization of computational screening models

for PK prediction can be found as one of the factors that reduced the frequency of PK-related attrition (Cook et al., 2014).

In silico ADME predictions have become ubiquitous as an important decision-making tool in the pharmaceutical industry (Kharkar, 2010), which is demonstrated with examples such as the in silico screen for efflux prediction that Eli Lilly adopted in place of an in vitro screen (Desai et al., 2013) – which will be discussed in further detail in section 2.8. Still, despite all improvement achieved throughout the last decades, attention should be drawn to the fact that, in both studies (Cook et al., 2014, Waring et al., 2015), the rate of failure due to PK/PD issues is significantly lower in preclinical stages (3%) than it is in phase I (15% and 10%). This is important to address as the cost of the withdrawal of a candidate drug increases exponentially as one moves further along the development pipeline (Wishart, 2007). Clearly, there is still room for improvement, as the most problematic PK cases still go unnoticed from preclinical studies into phase I.

The importance of addressing and optimizing ADME in early stages is not only justified by the reduced likelihood of attrition due to problematic PK properties, but it also helps directing animal testing for safety and efficacy as it provides some degree of information that will allow better dosing decisions in animal studies. This possibly avoids common problems associated with higher dosage administrations such as solubility issues or other formulation challenges (Peakman et al., 2015).

1.2. Motivation for Exploring Distribution through in Silico Modelling

As established earlier, PK has a very important role in determining failure during drug development. In particular, distribution (addressed as volume of distribution) has, alongside clearance, a paramount role in determining the duration of action, as the two relate directly to the elimination half-life, which is used for dose optimization and the establishment of appropriate dose regimens (Smith et al., 2015). While early stage in vitro automated assays can be performed to gauge distribution by testing the extent of interaction with different transporters, which can be used to infer an estimate for the extent of distribution, estimation of volume of distribution (Vd) itself still relies mainly on allometric scaling from animal data (Hop, 2015). Retrospectively looking at the performance of allometric estimation of Vd (corrected for animal plasma binding) showed 69% of predictions were within a two-fold deviation from clinical data (Smith and Baillie, 2015). However, this entails that protein binding assays have to be performed as well, since not accounting for this interspecies difference increases the error. Allometric scaling is also associated with other shortcomings

that arise from other sources of interspecies differences, which render it somewhat unpredictable. This will be discussed further in section 1.9.

Something as simple as low Vd can render unfeasible the use of a highly active, non-toxic drug due to its inability to properly reach the tissue in sufficient amount to elicit the desired response. Properties like this are often not discovered until human trials, which means that any drug withdrawals are extremely expensive for the company. As established earlier, in silico characterization of PK properties had a central role in reducing clinical stage attrition in the last decade and, as a result, computational models are a promising and inexpensive tool to minimize late drug attrition rate due to an unwanted distribution profile, as well as providing early information that guides dosage scheme design for first in-human trials. Among these, quantitative structure-activity (or property) relationships (QSAR or QSPR) have long been implemented in the drug discovery and development process. QSAR models of Vd have been able to achieve a level of predictive performance close to that of animal models (i.e. 2-3 fold error) (Gleeson et al., 2011), which further supports computer models as a competitive screening tool.

Because, as established earlier, there is still a significant portion of poor-PK candidates that are being approved to proceed into clinical trials, there is still a need to improve the capability of the in silico filters to detect such candidates early in the screening process. As a result, the work presented in this thesis consists mainly of an effort to provide new sources of model features i.e. molecular descriptors that might help elucidate the physiological processes that rule distribution, and consequently help modelling this endpoint.

1.3. A Brief Context in Pharmacokinetics

Pharmacokinetics can be defined as the study of the course of a compound (typically a xenobiotic) through the body upon administration. This course is composed of four main components, namely Absorption, Distribution, Metabolism and Excretion, which are usually referred to as ADME. Even though distinct, these can occur simultaneously and are interrelated at many levels (Fan and de Lannoy, 2014).

Upon being administered orally, a compound progresses through the lumen of the gastrointestinal tract where it is absorbed through the epithelial lining. The absorbed fraction will first go through the liver (via the portal vein), where it is amenable to undergo first-pass metabolism and biliary excretion, and the remaining unchanged compound will then reach systemic circulation where it is distributed to various tissues. At the same time, as the

compound is carried by blood perfusion, it can reach metabolizing organs, such as the liver, where it is transformed by different enzymes such as the CYP450 members. Besides the liver, many other tissues are metabolically active (e.g., lungs, skin, brain, etc) (Curry and Whelpton, 2017, Fan and de Lannoy, 2014).

Both the metabolic derivatives and the unchanged portion of the drug can be excreted as they reach various possible excretory and metabolizing organs, from which the liver (through biliary excretion) and kidneys are typically the most influential organs at this stage (Fan and de Lannoy, 2014).

This process, schematized in Figure 1.1, controls the duration of residence and the amount of drug that reaches the desired site of action, as well as the undesired locations that lead to toxicity.

Other routes of administration exist, such as intravenous, intraperitoneal, or subcutaneous, among others, however the process explained above remains unchanged from the moment a drug reaches circulation. The only major changing factors between administrations are the amount and speed at which a drug reaches systemic circulation.

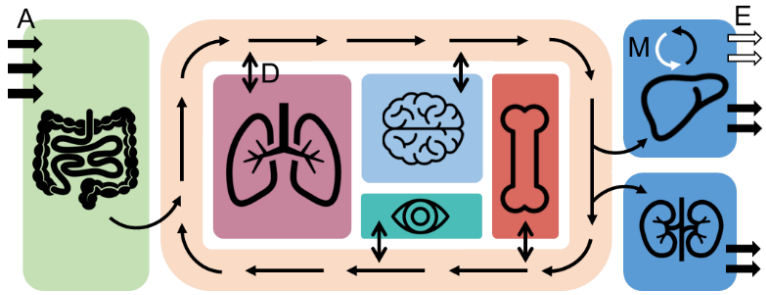


Figure 1.1. Process flow of Absorption (A), Distribution (D), Metabolism (M) and Excretion (E), encompassing the different processes through which an orally administered compound is submitted. This process is named ADME.

1.4. The Drug Distribution Process

Upon reaching systemic circulation (either through absorption or intravenous administration), a compound will be quickly distributed to highly perfused organs such as the kidney and the liver. This corresponds to the central compartment, where an equilibrium with systemic circulation occurs quasi-instantaneously. Additionally, the drug can be distributed to poorly perfused organs, and a second equilibrium with systemic circulation will occur. This corresponds to the peripheral compartment. This process is represented in Figure 1.2., and is called the two-compartment model. The ability to distribute to peripheral

tissues determines the extent of distribution (Wallace et al., 2011) and, overall, how a compound distributes across the body is a key determinant of its safety and efficacy (Gleeson et al., 2011).

Some of the main chemical determinants of the pharmacokinetics of a drug in general, and distribution in particular, are lipid solubility, molecular weight (Wallace et al., 2011) and ionisation state (which also affects the lipid solubility). Besides chemical features, there are physiological factors that drive distribution, such as blood flow rate, active transport and binding to plasma proteins or tissue structures (Wallace et al., 2011).

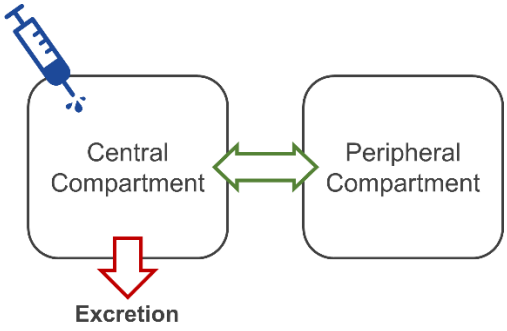


Figure 1.2. Two-compartment distribution model that accounts for tissue partition.

A measure used to represent the extent of distribution of a compound across tissues is volume of distribution, Vd, which simply shows how a drug dose relates to measured systemic concentration, as shown in Equation 1.1.

$$V_d = \frac{\text{Dose}}{C_{plasma}} \qquad \text{(Eq. 1.1.)}$$

Dose represents the mass of compound that effectively reaches the systemic circulation (bioavailable dose), and C_{plasma} represents the resulting concentration achieved in plasma upon administration. As a result, Vd does not represent a real physiological volume, however it may (but not necessarily does) reflect the relative ability of a drug to reach or accumulate at different tissue locations. Volumes of 3-5 L indicate that the compound is mostly limited to the intravascular space, whereas volumes of 30-50 L indicate an ample distribution throughout the total body water and, hence, the ability to likely reach the intracellular space. It is, however, possible to reach a Vd larger than total body water, through the occurrence of tissue partition phenomena (Wallace et al., 2011, Smith et al., 2015). Here, the displacement of a drug to tissues and its accumulation at certain tissues leads to a decreased plasma concentration, which is simply perceived as increased dilution of the drug in body water, as represented in Figure 1.3 (Wallace et al., 2011, Holford and Yim, 2016).

There are different volume terms derived from different stages of the pharmacokinetics curve of concentration versus time. However, the Volume of distribution at steady state (Vss) is the most used parameter as it is estimated by non-compartmental techniques (there is no underlying assumptions for its calculation) and it represents the apparent distribution in steady state conditions (stable equilibrium between the rates of input and output of a drug). As a consequence, this is the most appropriate value (and the most straightforward to obtain) for use in drug design (Smith et al., 2015).

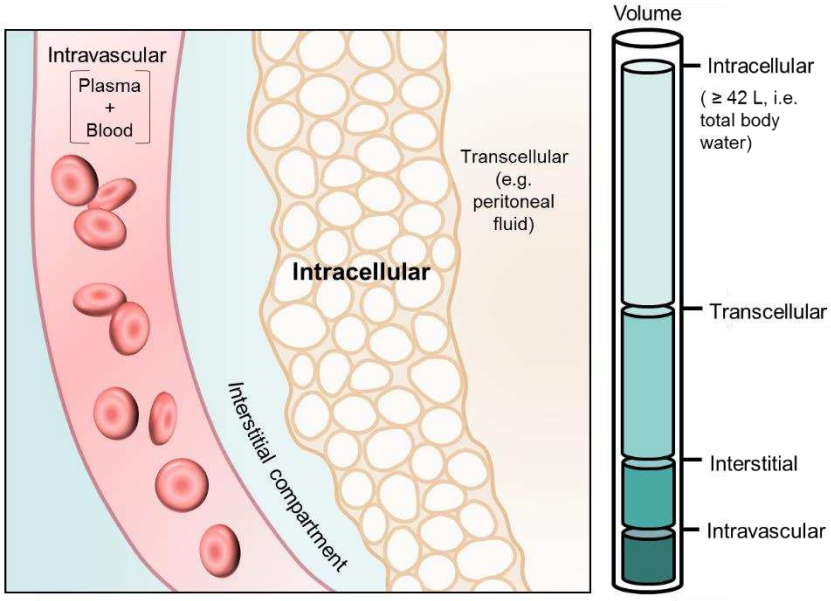


Figure 1.3. Representation of the relationship between observed values of Vss and their respective distribution into the different tissue spaces.

Vss is determined by the net contribution of plasma and all available compartments (i.e. tissues) to which a drug can distribute, and their respective volumes. In addition to this, the relative ratio of binding in plasma and in tissue defines the degree to which each of these compartments is occupied. As the bound fraction of drug is in equilibrium with the respective unbound fraction in both sides of the membrane, as shown in Figure 1.4, the latter can be used to represent binding extent as these are easier to measure.

Following this, Vss can be expressed as shown in Equation 1.2, where Vp represents the volume of plasma, Vt is the volume of any given tissue, and fu_p and fu_t are the fractions of unbound drug in plasma and tissue, respectively.

$$Vss = Vp + \sum \left(Vt \times \frac{fu_p}{fu_t} \right) \quad \text{(Eq. 1.2)}$$

As Equation 1.2 entails, binding to plasma proteins and to tissue structures is a major modulator of Vss. Besides this, there are a number of determinants, of either

physicochemical or physiological nature, that modulate the distribution process and affect the measured Vss, which are discussed next.

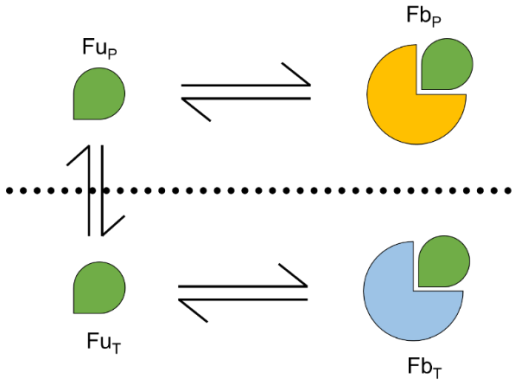


Figure 1.4. Membrane partition equilibrium. Only the fraction of free drug, Fu, will be able to cross the biological membrane, and this is determined by both the extent of binding to the tissue structures (Fb_T) and to the plasma proteins (Fb_P).

1.5. Physicochemical Determinants of Distribution

Given Vss is the result of the extent of partition across biological membranes into different tissue compartments throughout the body, one can establish that physicochemical factors that affect permeation will affect Vss. According to the Fick's law of diffusion, it has been established that only the unbound portion of a compound is free to undergo passive permeation. As a result, the acidic/basic character of a compound is one of the key factors of Vss as it determines the formal charge in physiological pH. According to their ionization state, compounds will interact with plasma proteins and tissue structures differently. For instance, acid species are more prone to bind (very strongly) to albumin, whereas basic species will preferably interact with α-1 acid glycoprotein as well as albumin (Smith et al., 2015, Curry and Whelpton, 2017). The ionization state also drives tissue binding as basic compounds, for example, are more prone to interact with acidic phospholipids in the plasma membrane and hence, more prone to tissue partition. In fact, the level of phosphatidylcholine in different tissues correlates with the relative levels of tissue partition of several positively charged compounds in physiological conditions. Lipophilicity determines the binding affinity to compounds that are amenable to interact with plasma proteins, and additionally it determines the extent of tissue and/or membrane binding of charged and neutral compounds (Smith et al., 2015).

As a result of this, basic compounds have a tendency to be extracted from systemic circulation and are normally associated with relatively larger Vss, whereas acids will tend to remain in the intravascular space or extravascular water as these have a high amount of albumin (Smith et al., 2015).

However, as noted by Smith (Smith et al., 2015), small Vss values do not necessarily imply an inability to partition into the site of action in a tissue, and the previous physicochemical considerations are general observations that do not always apply. When considering the unbound fraction only, some acidic compounds are able to partition into tissues at the same ratio as neutral or basic compounds do, despite large differences in Vss.

1.6. Physiological Determinants of Distribution

1.6.1. Tissue and Plasma Protein Binding

One of the most important physiological determinants of Vss is binding to structures in or outside tissues, and this factor is tied to the electrolyte type of the drug, as established earlier. A high degree of binding to tissue and low binding in plasma will generate increased Vss (given $Fu_P/Fu_T > 1$) (Smith et al., 2015, Curry and Whelpton, 2017). In the tissue, different compounds are able to bind to different macromolecules such as phospholipids, DNA, and carbonic anhydrase or acetylcholinesterase in red blood cells (Curry and Whelpton, 2017). In plasma, binding can occur to plasma proteins such as albumin and α-1-acid glycoprotein mentioned already, as well as (α-, β- and γ-) globulins and lipoproteins (Yanni, 2015, Smith et al., 2015).

Binding and its resulting entrapment in lysosomes has very marked effects on distribution, leading to very large observed Vss values. This phenomenon is called drug-induced Phospholipidosis (PL) and will be explored in more detail later in this chapter.

1.6.2. Blood Perfusion and Transfer across Membranes

Distribution is directly controlled by blood flow, as this controls the rate at which compounds are brought in contact with any given tissue. As mentioned earlier in this chapter, high perfusion (occurring, for example, in the liver), leads to quicker equilibration between tissue and plasma and larger amounts of drug will be partitioned into highly perfused tissues. On the other hand, poorly perfused tissues, such as the adipose tissue, experience slow

equilibration with plasma and receive a lower amount of drug (Yanni, 2015). A summary of the main organs and tissues and their blood flow is shown in Table 1.1.

Table 1.1. Standard blood flow in different human tissues, arranged in descending order (Yanni, 2015).

Tissues	Blood Flow (mL/min)
Lungs	5000
Liver	1350
Kidney	1100
Muscle	750
Brain	700
Skin	300
Bone	250
Heart	200
Fat	200
Spleen	77
Thyroid	50

In order for a compound to reach a tissue compartment it has to be able to cross its biological membranes using one or more of the available permeation routes such as paracellular passive diffusion (through pores between cells), transcellular passive diffusion (through the cell membrane's lipid bilayer), protein-mediated transport or endocytosis (entry through the formation of a vesicle). Consequently the structure of the epithelia will pose different levels of hindrance to permeation (Fan and de Lannoy, 2014, Curry and Whelpton, 2017, Yanni, 2015). Locations such as the liver (Fan and de Lannoy, 2014, Curry and Whelpton, 2017) or the brain (Cipolla, 2009) are formed with tight junctions which limit paracellular diffusion to relatively small molecules (up to 200 g/mol). Additionally, the presence or absence of fenestrations in peripheral capillaries that serve different tissues can change the access of a drug to a tissue (Curry and Whelpton, 2017).

1.6.3. pH and Electrochemical Gradients

The pH gradient observed between different compartments is also a driver of partition, following the general principle that unionized species encounter less hindrance when traversing biological barriers than their ionized counterparts (Fan and de Lannoy, 2014). A pH of 7.4 in the extracellular medium is, for example, slightly larger than the intracellular pH (~7.2), which is, in turn, much larger than the lysosomal compartment pH (~ 5) (Casey et

al., 2010). Variable pH determines the equilibrium between ionized and non-ionized species, following the Henderson-Hasselbalch equation (Equations 1.3 and 1.4).

$$\mathrm{pKa\ (acid)} = \mathrm{pH} + \log\frac{[\mathrm{non-ionized}]}{[\mathrm{ionized}]} \quad \text{(Eq. 1.3)}$$

$$\mathrm{pKa\ (base)} = \mathrm{pH} + \log\frac{[\mathrm{ionized}]}{[\mathrm{non-ionized}]} \quad \text{(Eq. 1.4)}$$

As a result, depending on the compound's pKa and its electrolytic nature, it preferentially distributes to different locations (Smith et al., 2015, Curry and Whelpton, 2017). Acids tend to concentrate in the more basic compartments, and bases tend to passively diffuse towards more acidic environments (Curry and Whelpton, 2017, Yanni, 2015). This is theoretically valid, however actual concentration gradients between compartments might be smaller given the constant flow of fluids between both places (Yanni, 2015). Similarly to pH, electrochemical gradients also drive partition across compartments. Electrochemically-driven partition occurs, for example, in mitochondria where a large potential difference exists, while pH-driven partition is prone to occur, for example, in lysosomes (Smith et al., 2015).

1.6.4. Membrane Transporters

Another modulator of distribution is protein-mediated transport, which can deplete or concentrate drug at any given compartment (Smith et al., 2015). Drugs that undergo uptake by membrane transporters will be displaced to the tissue (intracellular or interstitial space), reaching larger tissue-to-plasma ratios, whereas efflux will displace drugs to the plasma, leading to smaller tissue-to-plasma ratios (Wagner et al., 2016). Even though the effect of transporters is often deemed negligible towards Vss when compared to the effect of the unbound fraction (Smith et al., 2015), there are several examples of drugs (e.g. antiretrovirals, antihyperglycemics, hepatitis antivirals, guanethidine, sympathomimetics or paraquat) in the literature that are highly concentrated in tissue spaces due to transporter uptake (Wagner et al., 2016, Curry and Whelpton, 2017).

Additionally, transporter-mediated partition, as well as pH- and electrochemically-driven partition have important implications on the pharmacology and toxicology of drugs, as these outcomes are dependent on local distribution effects (i.e. accumulation on the site of action or on undesired cellular sites). This can be valid even when no impact over global Vss is visible (Smith et al., 2015).

There is a large amount of evidence of transporters, such as the P-glycoprotein (P-gp), that directly affect ADME as well as safety and, as a result, P-gp efflux assessment has become mandatory in all drug development campaigns. P-gp is part of a large family of efflux transporters called the ATP-Binding Cassette (ABC), which has an important role in ADME. Another key family is the Solute Carrier (SLC) transporters, which mediate uptake. Both families will be further described in the next section (Yanni, 2015).

1.7. Membrane Transporters and Their Role in Drug Distribution.

PK plays a major role in the success of a drug candidate to meet desirable properties during drug discovery and development. PK properties such as permeability, oral bioavailability, half-life and drug-drug interactions may all be affected by the binding of drug candidates to transporter proteins (Giacomini et al., 2010, Ballard et al., 2012). The FDA currently recognises the importance of transporter proteins in modifying drug exposure levels and as a source of potential drug-drug and drug-food interactions (FDA, 2012). There are mainly two super-families of transporters that are targeted in pre-clinical studies: The ATP-Binding Cassette (ABCs) and the Solute Carrier (SLCs) transporters. They are generally associated with clinically relevant impact on the efficacy, adverse effects and drug disposition profile of drugs, which derives from their important role as determinants of tissue access (Kharkar, 2010, Chu et al., 2013, Giacomini et al., 2010). As a result, ABCs and SLCs are the primary research focus in drug development (Wang et al., 2015) and the potential for efflux and uptake mediated by these transporters is typically studied during the preclinical stage to better propose clinical studies to address the precise impact of a given protein or group of proteins (Giacomini et al., 2010).

There are more than 400 transporters in these two superfamilies identified in the human genome. To this date many of those have been cloned and characterized. In drug development the focus is turned to transporters expressed in the liver, the kidney, intestine, and in the blood-brain barrier endothelium (Giacomini et al., 2010).

1.7.1. ATP-Binding Cassette (ABC) Transporters

The ABC family is composed (in humans) of 48 membrane transporters that are grouped in seven families from ABCA through ABCG (or ABC1, MDR/TAP, MRP, ALD, OABP, GCN20 and White, respectively). They show, in general, close homology over two ATP-binding domains and 12 putative transmembrane domains implicated in the efflux of xeno- and

endobiotics (Locher, 2016, Lai, 2013a), however this structure is not strictly observed in some cases, as will be pointed out later in this section. Most of the 48 ABCs are thought to mediate efflux, and in some cases they exhibit an outstanding substrate polyspecificity (Locher, 2016). The structures of the various transporters differ considerably but all of them have two separate domains: (1) the hydrophobic transmembrane domain (TMD) (or membrane-spanning domain, MSD), that has specific drug-binding sites; and (2) the nucleoside-binding domain (NBD). Although always present, different arrangements of both domains are found in the ABC proteins. The classical ABC transporter contains two domains of each type; however, they can also be formed by the dimerization of two half ABC proteins (like in the case of BCRP), or by two NBDs and three MSDs (like with MRP2) (Lai, 2013a).

Although always present, TMDs have diverse structures and are characterized by alpha-helices included in the bilayer (normally 12 alpha-helices, 6 per monomer). The alpha-helices recognize several substrates, which prompts reversible conformational changes that allows the crossing of the membrane (Lai, 2013a).

The name of this family of proteins – ABC – derives from the fact that it has a highly conserved ATP-binding cassette sequence motif. The ATP-binding domain is found in the cytoplasmic side, and it contains a signature C motif, specific to each family member, as well as two sequences – Walker A and B – present in all ABC transporters. These are highly conserved motifs where ATP is hydrolysed. The ABC transporters pump the substrates from the cytoplasm into the extracellular medium against their gradient by utilizing the energy released from ATP hydrolysis (Ford et al., 2010, Locher, 2016, Lai, 2013a). They are also responsible for transporting compounds from organelles into the cytoplasm, and some ABC members act as ion channels (Ford et al., 2010, Locher, 2016).

The currently established mechanism of efflux consists of an ATP-fuelled shift between an inner-facing “V”-shaped position that binds any given substrate and an outer-facing position (where the in-facing “V” gets inverted) that releases it into the extracellular medium (Ferreira et al., 2015, Locher, 2016). The inner-facing position corresponds to a high-affinity binding state that suffers conformational modifications induced by substrate binding and ATP hydrolysis, which leads to the adoption of a low-affinity conformation (the outer facing position) that releases the bound compound into the extracellular medium (Ferreira et al., 2015, Locher, 2016).

The binding sites are hypothetically accessible to the ligands from the inner core of the plasma membrane, or directly from the cytoplasm. The former would be expected to apply to strongly lipophilic compounds, which would tend to diffuse into the inner core of the membrane and be transferred from there into the binding pocket (Locher, 2016).

ABCs transport a wide variety of endogenous and exogenous compounds, which range from ions to macromolecules, via an ATP-dependent mechanism (Pinto et al., 2014, Marquez and Bambeke, 2011). These transporters are highly expressed in a variety of tissues, some of which are some important distribution barriers that are associated with drug absorption and distribution impairment (Szakács et al., 2008).

Efflux is one of the biggest challenges in pharmacotherapy and many ABC transporters have been demonstrated to have clinical relevance (Locher, 2016). The ability of a tissue to remove a compound from the intracellular compartment to the extracellular medium is used as a defence mechanism against metabolites and other noxious compounds, but at the same time it limits the bioavailability of various therapeutics in several tissues. This phenomenon hinders the success of a number of therapeutic regimens ranging from anti-cancer therapy to antibiotics. Often it is observed that various cases of therapeutic failure are associated with the overexpression of one or more efflux transporters, among which ABC proteins play a major role (Ferreira et al., 2015). Examples of ABCs that are strongly associated with multidrug resistance are Breast Cancer Resistance Protein (BCRP1 or BRCP, ABCG2), P-glycoprotein (P-gp, MDR1, ABCB1), and the Multidrug Resistance-associated Proteins (MRP1-7, ABCC1-6 and 10) (Pinto et al., 2014, Marquez and Bambeke, 2011). ABC transporters have also been implicated in the pathophysiology of Alzheimer's disease (Cascorbi et al., 2013), diabetes and atherosclerosis (Allen et al., 2015).

P-glycoprotein (P-gp)

This transporter is found in the small intestine, blood-brain barrier, and in excretory cells such as kidney proximal tubule epithelial cells and hepatocytes, and as a consequence it is important in the control of CNS access, intestinal absorption, and in urinary and biliary excretion (Giacomini et al., 2010). This transporter has been intimately connected to multidrug resistance and the resulting cancer therapy failure (Alfarouk et al., 2015).

P-gp contains 1280 amino acids and a molecular weight of 170 kDa. It has two symmetrical halves linked by an (approximately) 75-amino acid linker, distributed in 12 transmembrane helices and two NBDs. Five substrate-binding sites have been identified up to date. In order to accommodate the ability to transport several structurally unrelated drugs, P-gp is proposed to function through an induced-fit mechanism in which there are changes to the transmembrane segments, although the clear transport mechanism is yet to be established.

So far three transport mechanisms have been proposed. Initially P-gp was though to act as a pore that allows direct passage from the cytoplasm to the exterior of the cell. Following this, it was proposed that this transporter would act as a flippase, transporting molecules

between both membrane leaflets (from the inner to the outer) and ultimately into the external medium. The currently accepted theory of transport mechanism is that P-gp attracts substrates trough hydrophobic interactions, acting as a hydrophobic "vacuum cleaner" (Lai, 2013b, Ferreira et al., 2015). This hypothesis is based on the notion that, once on the membrane, the hydrophobic region of a substrate is naturally more soluble in the hydrophobic inner leaflet, thus building up in the membrane's core. Being on this region allows it to diffuse laterally to the transporter's binding pocket also located on the inner core of the membrane. Substrates are thought to access the binding cavity through two entrance gates (between TM4 and 6; and between TM10 and 12). Once the substrate binds, this is thought to prompt P-gp to undergo conformational changes that open its cavity to the extracellular side of the membrane, and subsequent exit of the substrate (Ferreira et al., 2015).

The P-gp substrates' molecular weights range from 300 to 2000 Da, often being amphipathic or hydrophobic organic cations (Giacomini et al., 2010, Lai, 2013b). P-gp binds to a wide variety of structurally unrelated ligands, and this has been explained through ligand-inducing change in conformation – known as the induced fit hypothesis (Ford et al., 2010).

According to the plasma membrane composition, it exhibits different temperatures of transition between fluid liquid-crystalline and rigid gel phase (Ferreira et al., 2015). Following the hydrophobic vacuum cleaner hypothesis, as the compounds are extracted by P-gp from the inner core of the lipid bilayer, it is logical that membrane composition will directly affect efflux, given that the ability to partition into the membrane will depend on the membrane's constitution. Furthermore, the lipid environment surrounding P-gp also impacts its ATPase activity and its ability to bind to ATP (Ferreira et al., 2015). On the other hand, increasing amounts of P-gp on the plasma membrane have shown to decrease t_m. A causal relationship has been demonstrated between the alteration of membrane properties (i.e., fluidity and morphology) induced by a number of small molecules and the modulation of affinity to the membrane, P-gp expression, and P-gp location and function (translocation causing ATPase inhibition) (Ferreira et al., 2015).

Breast Cancer Resistance Protein (BCRP)

BCRP belonging to the ABC subfamily G, different from either the MRPs' or P-gp's family. This transporter shows a high degree of tissue expression overlap with MRP2 and P-gp, being expressed in the liver, placenta, kidney, brain and intestine, among others, which might indicate some level of cooperation or redundancy in tissue protection responses.

BCRP is able to efflux various structurally diverse compounds, and has been attributed to have some degree of substrate overlap with P-gp (Lai, 2013d). Its ligands include antivirals, tyrosine kinase inhibitors, HGM-CoA (3-hydroxy-3-methylglutaryl coenzyme A) reductase, antibiotics and flavonoids. BCRP is also involved in the transport of sulfate conjugated drugs, preferentially over glucuronide- or glutathione-conjugates.

BCRP is formed by 655 amino acids and six (predicted) transmembrane helices (differentiating if from the typical 12-transmembrane domain structure of the ABC family), having a weight of 72 kDa. At least two binding sites have been proposed, but how these formed in the functional protein is still unclear (Lai, 2013d, Giacomini et al., 2010). QSAR analysis suggests that N-C(Heterocyclic ring) seems to favor drug interaction with the transporter; also fused heterocyclic rings containing two substituents on a carbocyclic ring have been identified as patterns that promote molecular recognition (Giacomini et al., 2010).

Multidrug Resistance Proteins (MRPs)

In this work two particular MRPs have been addressed – MRP1 and MRP2 – so this section will focus mainly on these. Both MRP1 and MRP2 are composed of 6+6+5 helices, arranged respectively in three membrane-spanning domains, two nucleotide-binding domains and a linker. There is evidence indicating that there are at least two binding sites; one is proposed to be involved in direct binding and the other in allosteric binding, regulating the affinity to the former site (Lai, 2013c).

MRP2 is formed of 1545 amino acids and exists as a 190 kDa phosphoglycoprotein. It is typically found on major physiological barriers like the liver, kidney, intestine or the placenta. MRP2 has an important role in extracting endogenous metabolites, such as bile salts, conjugated bilirubin or conjugated drug metabolites, into the bile. As with BCRP and P-gp covered earlier, MRP2 transports a variety of structurally diverse drugs and their metabolites. Structural patterns that determine molecular recognition include two lipophilic regions (such as aromatic rings) and an anionic ionisable group. Accordingly, QSAR studies indicate lipophilicity, hydrogen bond elements, polarizability and aromaticity as critical for binding. MRP2 is thought to be the main transporter in the biliary and renal excretion of organic anions (both parent drugs and their metabolite conjugates) (Lai, 2013c).

MRP1 is expressed in a number of tissues as well and it plays a key role in transporting sulfate-, glucuronide- and glutathione-conjugated compounds (Pinto et al., 2014). In the liver, MRP2 and OATPs act synergistically as evidenced through the example of

pravastatin. This drug is uptaken into the liver by the latter, and afterwards excreted from the liver into the bile by the former (Lai, 2013c).

1.7.2. Solute Carriers (SLCs)

SLCs form the largest superfamily of transporters (composed of 456 members spread across 52 subfamilies), and the second-largest among membrane proteins (Cesar-Razquin et al., 2015). The proteins in this family contain multiple TMDs and mediate the crossing of membranes either against or with the concentration gradient. Besides occurring on cellular plasma membranes, SLC transporters are also found on organelle membranes (endoplasmic reticulum, mitochondria, Golgi apparatus, etc.). The members of this family were originally named according to the general form of “SLC” + #family + “A” to “E” specifying the subfamily; however, this nomenclature has changed and now there are different designations for the various proteins of this superfamily (Lai, 2013a). Members are allocated to the various families according to a minimum of 40% amino acid sequence similarity, and they are assigned to subfamilies with a minimum of 60% sequence identity (Hagenbuch and Stieger, 2013).

The SLCs can be passive transporters (uniporters), if transport is powered by differences in electrochemical potential and substrates are therefore moved down their concentration gradient, or alternatively they can be secondary active transporters (both symporters and antiporters), if substrates are transported against their electrochemical gradient by being paired with the transport of an ion down its concentration gradient. Antiporters specifically can be of two types: solute-cation or solute-solute transport (Lai, 2013a).

From a pharmaceutical standpoint, SLCs seem to have garnered great interest, since they have the potential to be used as facilitators of the delivery of drugs to their targets (e.g., PepT1 for gastrointestinal mucosa and blood brain barrier (BBB) crossing) (Lai, 2013a). Additionally, they are also of great interest as pharmacological targets. An example of such relevance can be seen with OATP1B1 (also known as SLCO1B1), which drives a preferential distribution of statins into the liver versus the muscle, hence allowing acceptable therapeutic index by enabling a relatively lower extent of statin-related myopathy. In a more extreme scenario, SLCs have been found to be the single driver of availability at the binding site (*e.g.*, YM155 is a cancer drug candidate which relies solely on SLC35F2 for entry into tumour tissue) (Winter et al., 2014). As a result of their role in mediating the distribution of drugs into or out of the site of action, SLCs are also associated with drug-drug and drug-nutrient interactions that are caused by competitive transport. Additionally, SLCs also play a key role in different physiological mechanisms, and roughly 190 mutated SLCs have been

implicated in disease states (Cesar-Razquin et al., 2015, Artursson et al., 2013, Ishikawa et al., 2016).

Despite this clear importance of SLCs in health-related research, in a recent review Cesar-Razquin et al (Cesar-Razquin et al., 2015) reported this superfamily as having the most skewed distribution of publications across its members, where a few SLC members hold the bulk of publications dedicated to the SLC superfamily, leaving most unexplored. It was suggested in this review that exploring this group of underexplored proteins can potentially uncover relationships amongst SLCs, which in turn may guide future experimental exploration as well as uncover interesting druggable properties that guide drug discovery (Cesar-Razquin et al., 2015).

Peptide Transporter 1 (PEPT1)

PEPT1 is primarily known as a key facilitator of peptides' (specifically di- and tripeptides) absorption; however, it also modulates the disposition of many xenobiotics, which will typically show some steric resemblance to PEPT1's natural substrates (i.e. peptidomimetics, e.g beta-lactam antibiotics) (Brandsch, 2013, Flaten et al., 2011). It is expressed in a variety of tissues.

PEPT1 is formed by two sets of six transmembrane alpha-helices, and substrate transport is fuelled by H^+ symport. One binding site has been identified, in a hydrophilic cavity, however this is not capable of accommodating larger compounds such as valacyclovir or tetrapeptides (Fowler et al., 2015, Flaten et al., 2011), which hints at the possibility of at least another binding site.

The transport mechanism has been proposed to follow a rocker-switch movement which has recently been expanded to a double scissors hypothesis, where two blades, one from each scissors, open in concert to each side of the membrane. The transporter alternates between outward and inward facing conformation, where it requires the apo conformation to switch into the outward conformation, and it needs to be bound to a substrate (holo conformation) and at least one proton to switch back into the inward conformation (Fowler et al., 2015).

Organic Cation Transporter 1 (OCT1)

OCT1 belongs to the SLC22A subfamily. This subfamily includes the OATs (significant human isoforms: URAT1, and OAT1-4 and 7) and the electrogenic OCTs (isoform 1-3). The

OCT1-3 are composed of 542-556 amino acids, with 12 predicted alpha-helical transmembrane domains (Lai, 2013f, Giacomini et al., 2010).

OCTs exhibit broad substrate specificity, but OCT1 typically transports type I cations (fixed charge) of hydrophilic nature and low molecular weight in a sodium-independent process. However, it is also capable of transporting anions (Lai, 2013f, Giacomini et al., 2010). Pharmacophore studies have concluded that hydrophobicity and positive charge are important elements for molecular recognitions by OCT1 (Lai, 2013f).

Despite the literature reporting OCT1 as primarily expressed in the liver, with minimal expression in other tissues (Lai, 2013f), this has been confirmed to be incorrect with recent direct protein quantification in a large number of tissues, reported in the human proteome atlas (Uhlén et al., 2015). Nonetheless, OCT1 plays an important role in facilitating hepatic excretion of compounds.

Organic Anion-Transporting Polypeptide (OATPs)

OATPs (or SLCOs) have a characteristic 12 transmembrane-helix structure and 5 extracellular loops. This family is found in different tissues including the BBB, intestine, liver and muscle, where they play a key role in modulating absorption and overall tissue access. Typical OATP substrates are amphipatic anions, however they can also be neutral or zwitterionic in nature, and they include compounds such as steroid conjugates, bile acids, oligopeptides, thyroid hormones, and various different drugs (Lai, 2013e, Giacomini et al., 2010). OATPs represent a high risk for drug-drug interactions and, as a result, both the American and European drug agencies (FDA and EMA) require an in vitro characterization of the interaction with OATP1B1 and OATP1B3 for every drug candidate eliminated hepatically, as these are they key modulators of biliary excretion (Lai, 2013e). OATP2B1 has a relatively narrower substrate specificity than the two former transporters (Lai, 2013e).

The OATPs are generally thought to have various binding sites. OATP1B1, for example, has been proposed to have at least two different binding sites, one with higher affinity and the other with lower affinity, and OATP2B1 has also been proposed to have multiple binding sites (Shirasaka et al., 2012, Tamai and Nakanishi, 2013).

The uptake mediated by several OATPs (e.g. OATP2B1) has been demonstrated to be promoted by acidic extracellular pH (Hagenbuch and Stieger, 2013) and it is thought to be coupled with the efflux of glutathione and glutathione conjugates. Co-transport of bicarbonate has also been proposed, however there is contradictory evidence that points

both to the enhancement and inhibition of uptake paired with bicarbonate efflux (Lai, 2013e, Giacomini et al., 2010).

1.8. Phospholipidosis and its Role in Drug Distribution

Lysosomal trapping plays a major part in tissue binding, therefore being a key driver of distribution, and it has gained increased importance with regard to distribution and toxicity (Logan et al., 2013). Lysosomal uptake of xenobiotics (also called lysosomotropism) can develop into a disorder where there is an excessive accumulation of phospholipids and xenobiotic in the tissue, called (drug-induced) phospholipidosis. These phospholipidosis-inducing xenobiotics are typically (but not necessarily) cationic amphiphilic drugs, which are prone to partitioning into the lysosome as, inside this compartment, the non-ionized form finds a more acidic pH which drives the protonation and subsequent conversion into the ionized species. As summarized in Figure 1.5, while the ionized fraction is unable to cross the membrane, the non-ionized fraction gets depleted with continued partition of neutral drug into the lysosome. As the concentration of the ionized base increases inside the lysosome, so does the medium pH. This means that all enzymatic processes that normally take place at pH 5 will be less favourable (Shayman and Abe, 2013, Smith, 2016). From this point there are several proposed mechanisms to explain the increase in phospholipid content inside the lysosome, with all theories indicating the impairment at some level of the lipid metabolism pathway, by interacting with either lipids or lipid-related enzymes (Shayman and Abe, 2013).

The affected cells exhibit myeloid or lamelar bodies (membranous, concentric structures consisting of deposits of undegraded lipids) in the cytoplasm and an overall “foamy” appearance. This phenotype is only confirmed with transmission electron microscopy, and has been reported in different types of cells (alveolar and lymphoid tissue macrophages, hepatocytes, renal epithelial cells, neurons, or bile canalicular cells, among others) (Shayman and Abe, 2013, Reasor et al., 2006, Anderson and Borlak, 2006). Phospholipidosis is associated with impaired lysosomal protein degradation, compromised ability for endocytosis, induced formation of free radicals as well as compromised immune response. Alternatively to being identified as a toxicity mechanism, due to the occurrence of previously listed outcomes, there is also a theory that phospholipidosis is an adaptative defense mechanism, and is not intrinsically toxic. This theory claims that drugs that are potentially toxic to other locations inside the cell are sequestered into lysosomes and

excreted bound to lamellar phospholipid-rich bodies. The lamellar bodies are then cleared upon secretion by macrophages (Shayman and Abe, 2013, Anderson and Borlak, 2006).

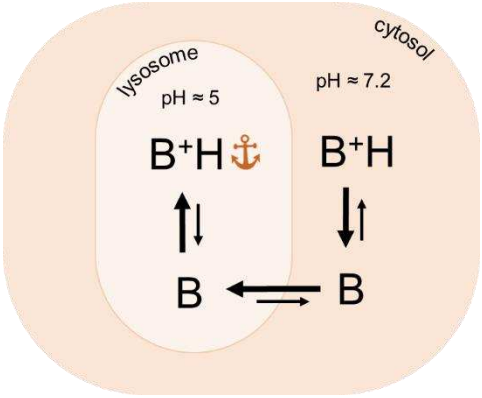


Figure 1.5. Representation of the ion partition equilibrium that drives basic compounds to be entrapped in the lysosome.

Due to this concentration effect that characterizes phospholipidosis, there are important implications to Vss and tissue distribution patterns. There is evidence of drugs, such as quinacrine, which exhibit intracellular concentrations several hundred times larger than in the extracellular medium, and lysosomotropic compounds are frequently associated with unusually large Vss values (Logan et al., 2013). It is worth noting that interspecies extrapolation is unreliable, and even with human cell culture phenotypical evidence may take days and even months to develop, which has prompted the interest in in silico approaches to predict phospholipidosis (Reasor et al., 2006).

Interestingly, both the concentration of drugs in the lysosome and their elimination (reversal of lamellar bodies) from within the cell are associated with transporters (Reasor et al., 2006, Shayman and Abe, 2013).

1.9. Experimental Determination of Drug Distribution Parameters.

In vivo determination of Vss typically relies on the collection of systemic concentrations (typically in plasma but also possible in blood) of a drug over a period of time after intravenous administration. The value of Vss can be derived from the administered dose and its resulting area under the curve (Fan and de Lannoy, 2014).

Drug distribution is most commonly investigated with in vivo animal models in preclinical stages (Yanni, 2015). However, relying on animals to extrapolate human PK has been demonstrated to be complex as well as unreliable (Tsaioun et al., 2016). Various OATPs,

for example, are found in preclinical species but are absent in humans, which creates possible issues in extrapolating findings from preclinical stages into human trials (Lai, 2013e). Due to the need to improve both throughput capacity and predictive power, as well as the need to reduce animal testing, it has become common to also use in vitro and in silico models to predict distribution during drug discovery. These played a vital role in improving the efficiency in the pharmaceutical industry, as discussed in the first section of the Introduction (Yanni, 2015, Tsaioun et al., 2016).

In a typical drug discovery workflow, plasma protein binding is measured during hit-to-lead screening, followed by lead optimization where tissue distribution is studied in one rodent species. Lastly, prior to administration in humans in Phase I, for advanced lead optimization, a candidate's distribution is assessed on one non-rodent species (dogs or non-human primates) and on one rodent species (Tsaioun et al., 2016, Zhang et al., 2012).

For plasma protein binding determination, a number of assays are used by the pharmaceutical industry such as equilibrium dialysis and ultrafiltration, which have high throughput (allowing 96-well plate testing). Other used methods include ultracentrifugation and chromatographic separation, and some less used methods such as exclusion chromatography, dynamic dialysis and circular dichroism (Yanni, 2015).

Animal studies for the assessment of the distribution of drug candidates are done through the administration of radiolabelled drug. Such studies include mass balance and quantitative whole-body autoradiography. In these, the labelled drug is administered orally or intravenously to the animal. In mass balance studies blood samples are collected over time, and tissues, urine and faeces are collected at specific times to determine whole-body distribution. The second technique, autoradiography, allows mapping distribution across all tissues and organs with time, however with the caveat of requiring a large amount of animals for each time point, and the inability to distinguish between metabolites and the parent compound (Yanni, 2015).

Radiolabelled studies are also performed in humans, namely positron emission tomography and magnetic resonance imaging allow monitoring the distribution of a drug into different organs. However, both in animals and humans, radiolabelled studies have the disadvantage of presenting possible safety risks associated with radioactivity exposure. Additionally, the cost and labour-intensive work associated with producing radiolabelled compounds makes this a prohibitive approach to investigate a drug's distribution, especially in humans (Yanni, 2015).

Drug distribution is also assessed by carrying out in vitro transporter studies. The standard parameter used for classifying a compound as a substrate or non-substrate of a given

transporter is the observed efflux/uptake ratio between the basal-to-apical partition and the apical-to-basal partition across a cell culture monolayer. This will typically be used to monitor the uptake by a given transporter of interest. However, it should be noted that if a cell line which expresses a variety of transporters, such as Caco-2, is used it is likely that the observed efflux ratio is a result of several transport routes (Crivori et al., 2006).

Some in vitro assays to carry out transport interaction studies include membrane vesicle assays where cells transfected with ABC transporters (or membrane vesicles obtained from transporter-expressing organs) are incubated with and without ATP to investigate dependencies between ATP content and permeation (Yanni, 2015, Tsaioun and Kates, 2012, Zhang et al., 2012). As a refinement of the vesicle assays, oocyte transport expression systems are used. These have also been reported to allow more precise assessment of transport by a specific transporter than cell lines (Shirasaka et al., 2012), which are another model to carry out transported impact studies.

Cultured cell models typically use Caco-2 or MDCK-transfected cells to study both efflux and uptake. Caco-2 cells have the advantage of differentiating into polarized enterocytes, acquiring tight-junctions and expressing transporters in a way that resembles the human epithelium. These are commonly employed in bi-directional permeability assays. As for transfected cell lines, such as MDCK, CHO, HEK293 or LLC-PK1, they have the advantage of allowing to study isolated (typically over-expressed) transporters (Yanni, 2015, Tsaioun and Kates, 2012, Zhang et al., 2012).

The impact of transporters on distribution into specific tissues can also be assessed by using primary cell lines such as hepatocytes, proximal tubular cells, or co-cultures of glial cells and brain capillary endothelial cells (Yanni, 2015, Tsaioun and Kates, 2012). However primary cell culture is more challenging from the technical point of view. For more specific analyses, primary cells collected from genetically polymorphic human subjects, or from transporter-deficient animals, are used. To address tissues such as the brain, MDCK cells expressing P-gp are commonly used as they form tight junctions also found in the blood-brain barrier (Yanni, 2015, Tsaioun and Kates, 2012).

In situ organ perfusion models are the closest surrogate of in vivo drug transport physiological processes, with the liver perfusion being the most used model. After a drug is perfused through the organ, it is possible to determine the amount of uptaken drug among other outcomes (Zhang et al., 2012).

1.10. Summary

One of the main standing issues in pharmaceutical industry is the still high attrition rates. However, important strides have been made towards managing this issue among which is the strategic decision to address the ADME profile early on. This was first done with in vitro models and followed by in silico models. Using computational approaches has proven to be a very useful and inexpensive tool to helping flag potential ADME-related issues. The ADME profile consists of the four main processes that administered compound undergo, namely Absorption, Distribution, Metabolism and Excretion, and this thesis is focuses on the distribution component.

Distribution is the group of phenomena that drive the partition of a substance between the various compartments in an organism and, as such, it determines the extent to which a drug reaches its site of action or a site of toxicity. Distribution is typically represented as a measure of volume called volume of distribution (Vd), and it is most reliably determined in steady state conditions (Vss). Many physiological and physicochemical factors play a role in modulating distribution, among which transporters have a large and complex impact. Two families of transporters are of particular importance: the ABCs and the SLCs. Another very important factor with drastic impact on distribution, which has been relatively underexplored is phospholipidosis.

Current experimental methods to characterise distribution range from simple in vitro assays to cell-based and animal assays. These present some limitations related to the fact that these are typically low throughput, expensive and resource intensive, which prompted the exploration of computational alternatives to predict distribution, such QSAR models, which derive a link between chemistry and a biological outcome.

2. Introduction Part II: QSAR modelling - Theory and Applications to Drug Distribution

2.1. Introduction to QSAR modelling

In the field of drug discovery it has become crucial to understand and establish a correspondence between a given activity or property of interest and the chemical structure. This field of research is generally named as Quantitative Structure-Activity Relationship (QSAR), even though it can be used to investigate not only the relationship between structure and activity, but also property, toxicology or selectivity. However, given that the methodologies of the four variants are the same, they are normally referred as QSAR, which will be done throughout this work (Gedeck et al., 2010).

The applicability of QSAR to any given problem relies on the premise that, in that particular problem, the variation in a measured property (e.g., binding, dissolution, inhibition, etc.) across a range of compounds is attributed (or at least correlated) to structural, chemical or physical variations (Goodarzi et al., 2013). In practice, a QSAR model is built by establishing a mathematical function that relates features (or descriptors) of compounds and their respective property readout. This function is typically produced by a machine learning algorithm (Roy et al., 2015).

The rationale behind having QSAR modelling as a competitive alternative to other available methods is two-fold. Firstly, as physicochemical features can be determined much more efficiently than the endpoint property of interest (determined in vitro or in vivo), computational predictive models offer a great reduction in cost, labour, and time (Goodarzi et al., 2013, Gedeck et al., 2010). To illustrate the difference in time-scale and cost between both scenarios, the patch-clamp technique used as an in vitro model of the hERG blockade consumes many research grade chemicals and requires one person for one work day, while a (previously trained) QSAR model to predict the same endpoint generally takes a few seconds and negligible effective cost (Gedeck et al., 2010). In fact, when analysing this scenario objectively, the first method is also a predictive model that has limited predictive power (Recanatini and Cavalli, 2008). The only exception to the short time required for in silico prediction is when high-level quantum mechanics is applied (Gedeck et al., 2010); however, precisely for this reason, these methods are more of an exception than the rule.

The second advantage of QSARs is that they are able to produce predictions on new, theoretical compounds, without requiring that these be synthesized. This means saving 1-2 weeks of synthesis in an optimistic scenario (Goodarzi et al., 2013, Gedeck et al., 2010).

To assure the quality of QSAR models, there are some general criteria that need to be followed during the construction of any QSAR model, and guidelines on this regard have been published (Tropsha and Golbraikh, 2010, Tropsha, 2010, Fourches et al., 2010). These guidelines will be briefly enumerated here, and further expanded in the next subsections. Firstly, the training set should be composed of sound experimental data, in order to avoid providing conflicting or fictitious patterns to the machine learning algorithm. This dataset should be annotated preferably with interpretable descriptors (Gedeck et al., 2010), and when a large pool of descriptors is available, feature selection must be carried out. This is one of the most important approaches to improve the predictive performance of a QSAR model (Goodarzi et al., 2013), and will be discussed further in a separate subsection. Once the dataset is ready for modelling, using reliable and robust algorithms is also important to assure that a useful QSAR is produced (Gedeck et al., 2010).

After the QSAR model has been built a comprehensive validation procedure is necessary to properly gauge its predictive performance. It has been demonstrated that the fit of the data used to build the model cannot be used as evidence of the model's predictive power or generalizability, and strategies like cross-validation and external-set testing are required (Yousefinejad and Hemmateenejad, 2015). It is also necessary to assess the model's confidence in outputting new predictions and this can be addressed through the characterization of the applicability domain. Another criterion that has been receiving increasing attention is the construction of interpretable models. This usually entails a power-to-interpretability trade-off, which is not straightforward to balance depending greatly on the expectations and needs that drive a given modelling task. However, considering that the real-world use of these tools needs regulatory and governmental approval, one could say that interpretability becomes increasingly more important, since it is translated into more confidence from the decision makers in a given predictive model (Gedeck et al., 2010).

2.2. Molecular Descriptors

Molecular descriptors are formal numerical representations of molecular structure, derived from a defined molecular representation using a specified algorithm (Danishuddin and Khan, 2016). Molecular descriptors play a key role in establishing statistical models of various endpoints of interest in health research and toxicology, among other areas.

Currently there are more than 5000 descriptors divided into different categories according to complexity (i.e. information content) of molecular representation (Consonni and Todeschini, 2010), and the information they encode typically depends on the algorithm used for their calculation as well as the kind of molecular representation. The simplest type of molecular descriptors – constitutional descriptors – characterizes compounds according to atom type of fragments in the molecule, as well as bulk physicochemical properties, such as the number of hydrogen acceptors. These descriptors do not account for molecular topology, hence they are not able to distinguish between isomers (Consonni and Todeschini, 2010, Danishuddin and Khan, 2016).

The second category in terms of complexity of molecular representation relies on the topological representation of molecules, and for this reason descriptors belonging to this group are named topological or 2D-descriptors. These descriptors take into account internal atomical arrangements and play a significant role in drug design, virtual screening, lead discovery and combinatorial library design, among others. Topological descriptors carry information about molecular shape, size, branching as well as heteroatomic and bond content. The calculation of these descriptors derives from a molecular graph rendition of a molecule (Consonni and Todeschini, 2010, Danishuddin and Khan, 2016).

At a higher level of complexity there are 3D (or geometrical) descriptors, which are derived from three-dimensional molecular conformation. Geometrical descriptors have high information content and are typically employed towards the discrimination between similar molecular structures and/or conformations, also encoding information on van der Waals areas across the molecular surface. As a direct consequence of this, they require geometry optimization, which makes them considerably more computationally expensive. Additionally, flexible molecules can be associated with several conformations, which entails an increasing complexity in finding a solution for the conformation state (Consonni and Todeschini, 2010, Danishuddin and Khan, 2016).

4D descriptors form the most complex type of descriptors as they result from the interaction energies between a probe and a grid-bound molecule (Consonni and Todeschini, 2010).

However, as pointed out by Consonni and Todeschini, descriptor complexity does not equate to its informative/predictive value, and this is determined on a case-by-case scenario, across the different types of endpoints to be modelled. For this reason, descriptors should be selected in a data-driven manner, through feature selection procedures, as explained next.

2.3. Feature Selection in QSAR

The new paradigm for data science has now shifted onto high dimensionality and/or large sample size. Highly dimensional datasets are typically associated to a high level of noise that is introduced by experimental error and by the fact that data comes from different sources (Tang et al., 2014). Further noise is introduced from the presence of irrelevant or redundant independent variables (normally named features), which can dilute meaningful patterns in the data. As a consequence, if used as they are, highly dimensional datasets are normally associated with training impairment (mostly due to overfitting that comes from chance correlations) in the construction of machine learning models – this is broadly referred to as "the curse of dimensionality". Therefore, whenever dealing with such datasets, the implementation of a pre-treatment routine prior to model training is of paramount importance (Goodarzi et al., 2013, Tang et al., 2014, Aggarwal, 2014). Dimensionality reduction is among the most used pre-treatment techniques for noise reduction through the removal of redundant and irrelevant variables. These techniques usually contribute to improved model generalizability and learning performance, as well as lower computational cost of training. Feature selection selects a small set of features that is able to maximize relevance and minimize redundancy with respect to the endpoint target of interest. Furthermore, the subset of selected features increases the model's interpretability (Goodarzi et al., 2013, Tang et al., 2014). In addition, normally a biological activity or property only requires a relatively small set of descriptors to be properly modelled into a QSAR (Goodarzi et al., 2013).

Feature selection algorithms can be broadly divided into filter, wrapper, and embedded methods (Bolón-Canedo et al., 2013, Saeys et al., 2007, Goodarzi et al., 2013), however only filter and wrapper methods are pre-processing methods. Filter methods score the features with respect to an intrinsic property of the data (e.g. the correlation between each feature and the class variable) and the highest scoring features are kept. Wrapper methods carry out a feature search that is guided by the performance of the machine learning algorithm to be used to build the QSAR model, so that the set of features that maximizes the cross-validation predictive performance is selected. This is carried out through an iterative process where a given candidate feature set, which is chosen by the searcher, is evaluated by the machine learning algorithm, which trains a tentative model with the candidate feature set. The performance obtained by this model is used to bias the direction of the search for the following iteration (Aggarwal, 2014). Lastly, embedded methods carry out feature selection during the training stage and are specific to the machine learning algorithms to which they are associated (Bolón-Canedo et al., 2013).

Comparing a filter and a wrapper approach, the latter is expected to yield more informative feature sets as it accounts for the bias in the learning algorithm, evaluating features in a context closer to the actual modelling task (Huang et al., 2007, Aggarwal, 2014). However, wrapper methods tend to be much more computationally expensive than filter methods, since the former need to run a machine learning algorithm many times, unlike the latter.

2.3.1. Genetic Algorithms

In a Genetic algorithms (GA) search, the algorithm finds a solution to a problem by applying the principle of natural evolution (survival of the fittest) to a population of chromosomes (candidate solutions represented as sets of features), which are submitted to genetic operators in an attempt to filter out non-critical information along the search process. Such genetic operators are crossover (the features in a pair of chromosomes are mutually exchanged) and mutation (a random turn on/off of one or a few single features in a chromosome) (Goodarzi et al., 2013).

In a genetic search the full list of features is randomly sampled into N subsets (with N determined by the user). This can be viewed as a random first set of N guesses of what a good feature set might be. These N subsets are then evaluated with a merit function and a percentage of the subsets are picked, from which subsets are paired according to their merit scores (higher is paired with higher) to mutually exchange a portion of their features (crossover operation). This is followed by random selection of features (under a pre-defined probability) to add or remove. At this point a new group of N child subsets are obtained, corresponding to the first generation (or first iteration) of the genetic search. The child subsets are evaluated with the merit function and the process repeats for a pre-defined number of generations. At the end of the process, a final group of N feature subsets is obtained, and the highest scoring subset is selected as the solution found by the search (Goldberg 1989).

In WEKA the merit function (Hall 1999) is defined by Equation 2.1:

$$\text{Merit} = \frac{\text{k} \times \text{mean}(\text{feature}-\text{class correlation})}{\sqrt{\text{k}+\text{k}(\text{k}-1)\times\text{mean}(\text{feature}-\text{feature correlation})}} \qquad \text{(Eq. 2.1)}$$

The advantage of using a GA as a feature selection algorithm is its ability to perform a global search (i.e., they are less likely to get trapped into local optima in the search space), thus tending to cope better with interaction between features. GAs are also capable of covering a large search space in a robust manner (Goodarzi et al., 2013, Tang et al., 2014). However,

a GA is rather prone to overfitting and, since it is a non-deterministic method, its success depends on the randomly generated initial population of chromosomes. To take into account this non-determinism, it is recommended to run repeated searches from different initial starting points in the dataset (Goodarzi et al., 2013).

2.3.2. Greedy Stepwise

Greedy Stepwise search (GS), or simply termed greedy search, performs a search through the space of features where it either starts with the empty feature set and tentatively adds new features, one at a time (forward search), or it starts with the full set of features and tentatively removes individual features (backward search). At each step, the best local decision is made regarding which feature to add or eliminate, respectively, such that changes maximize a merit score, defined by Equation 2.1. It is worth noting that no backtracking is carried out. As a result, with each new step, the search becomes increasingly limited, so GS guarantees to find a local optimal solution but not the global optimal one (Blum and Langley, 1997, Witten et al., 2011). The trade-off is that GS is robust against overfitting, and computationally inexpensive (Tang et al., 2014).

Greedy search strategies (within wrapper models) have been reported as very successful options for feature selection, since they are robust against overfitting and computationally advantageous (Tang et al., 2014).

2.3.3. ReliefF

Contrarily to some other feature selection algorithms, Relief algorithms (where ReliefF belongs) do not assume conditional independence of features. This makes ReliefF more appropriate to handle problems which entail considerable feature interaction. Additionally, ReliefF is efficient and aware of context information, being able to capture local dependencies which are normally missed by other methods. This is a robust algorithm which can deal with noisy and incomplete data, and has good ability to preserve sample similarity in a supervised learning context (Robnik-Šikonja and Kononenko, 2003, Bolón-Canedo et al., 2013, Zhao et al., 2013). Overall, ReliefF has shown excellent performance in real world machine learning applications (Zhao et al., 2013).

This algorithm evaluates features by iteratively selecting an instance at random and evaluating its k nearest neighbours of the same class (nearest hits) and its k nearest neighbours of a different class (nearest misses). The algorithm will reward features whose

values effectively discriminate the current instance's nearest hits from its nearest misses, by increasing their weight (Robnik-Šikonja and Kononenko, 2003).

2.3.4. Correlation-based Feature Selection

Correlation-based feature selection (CFS) is a filter algorithm that ranks features according to the correlation amongst themselves as well as their correlation with the dependent (class) variable. CFS searches for a feature set where features maximize the observed correlation with the dependent variable, while penalizing feature inter-correlation. The net effect of both these factors is quantified through a merit score (Bolón-Canedo et al., 2013).

The CFS algorithm works by first discretizing the numerical features in the training data, followed by an exhaustive calculation of feature-feature and feature-class correlations. This is paired with a user-selected searching algorithm (e.g. greedy search or genetic search) that searches through the feature space, and then the merit scores are calculated. The feature set that produces the highest merit during this search is kept to yield the final reduced-dimensionality dataset (Hall and Smith, 1999).

2.4. Machine Learning in QSAR: Regression and Classification

The machine learning methods used to build QSAR models are automated statistical tools that harness chemical and structural patterns towards describing a given target response. These are divided into two main groups of methods – Regression and Classification – which are employed to handle continuous (quantitative) and categorical (qualitative) response variables, respectively (Roy et al., 2015). Within both categories, the selection of the employed machine learning algorithm relies on two main aspects and their relative importance for the task at hand: interpretability and predictive power. Both aspects are important and offer different advantages, however they generally do not coexist in the same algorithm. As a result, one has to strike a balance between both aspects when modelling any given endpoint, which involves attributing priority to either interpretability or predictive power. This issue divides the machine learning community, and the QSAR community in particular (Fujita and Winkler, 2016). Whether emphasis is placed on either interpretability or predictive power depends on the purpose of the model's use. In cheminformatics, especially when modelling clinically relevant endpoints, the transparency of the model allows shedding light on the phenomenon captured by the dependent variable (Freitas, 2013).

2.4.1. Regression

The task of modelling a continuous response is called regression, and it can be divided into two types: linear and nonlinear. In linear regression there is a continuous response Y that occurs as a function of one or more independent variables (also called predictors or features) X in a linear manner. If there is more than one independent variables, the applied method is defined as multiple linear regression. These predictors are weighted with coefficients (a, b, … z), as shown below (James et al., 2013).

$$\mathrm{Y} = \mathrm{a} + \mathrm{b}\cdot\mathrm{X}_1 + ... + \mathrm{z}\cdot\mathrm{X}_\mathrm{n} \qquad (\mathrm{Eq.\ 2.2})$$

The coefficients (and predictors, if any feature selection procedure is used) are estimated in order to minimize the discrepancy (or residual) between observed (y_i) and predicted ($\hat{y}_i$) response, which is measured by the sum of squared residuals (James et al., 2013), as shown in the following equation, where N stands for the number of instances.

$$\sum_{\mathrm{i}=1}^{\mathrm{N}}(\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i})^2 \qquad (\mathrm{Eq.\ 2.3})$$

Equation 2.2 assumes a linear and additive nature between the response and the predictors, however this assumption can easily be violated if predictors show any sort of interaction amongst themselves. This can be overcome by allowing any number of predictors to be combined in one same additive term (James et al., 2013).

The linear regression described so far is produced by restricting the machine learning algorithm to learn a linear pattern between the various predictors and the response. Even though linear models are straightforward to implement and allow easy inference and interpretation, if a non-linear pattern exists between predictors and response, linear regression cannot appropriately model the problem at hand. In real-world data the assumption of linearity is, at best, an approximation and will very often not be applicable at all. There are several non-linear methods to address such scenario, namely polynomial regression, regression based on smoothing splines, generalized additive models, local regression and step functions (James et al., 2013).

2.4.2. Classification

The task of modelling a qualitative (also called categorical) response is called classification. Contrarily to regression, where the aim is to establish (through training) which numerical values the response Y takes throughout the span of one or more predictors, the main aim

in classification is to find numerical thresholds (called decision boundaries) in the predictor space that best separate the different categories or classes (values of the dependent variable), as represented in Figure 2.1. Such boundaries may be described linearly or non-linearly. Using the established boundaries, the built models predict the class of an instance, often through initially predicting the probability of each available class (James et al., 2013).

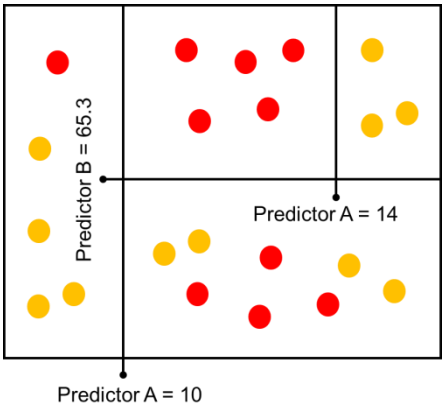


Figure 2.1. The square represents the full dataset available for training. The data is sorted using decision thresholds applied to Predictors A and B, which separate the data from two different classes (depicted in yellow and red, respectively).

Examples of classification algorithms are (linear and multiple) discriminant analysis, logistic regression, k-nearest neighbours, decision trees and other tree-based methods (such as random forests or boosted trees), support vector machines (SVMs), neural networks, and Naïve Bayes, among others (James et al., 2013, Madan et al., 2013, Aggarwal, 2014).

2.5. Machine Learning Algorithms

2.5.1. Decision Trees

Decision trees predict the value of the response variable, also called the output (which can be either continuous or categorical) by arranging data according to a range of successive partitions of the input space. This is done by carrying out a greedy search through all available descriptors (features), and determining which decision splits provide the greatest gain in each successive node, i.e. the most accurate partition of instances according to their correct class (for classification) or numerical partition of the response (regression). This process is recursively applied until all instances in each final node (leaf node) are of the same class or same subportion of the response range. During this tree growth phase, pre-

pruning can be used consisting of some termination criterion (e.g. the minimum number of instances in a node) that is meant to control tree complexity. After the tree is fully grown, it undergoes pruning in the post-pruning phase to avoid overfitting and allow better generalizability. There are different pruning techniques, from which three are available in WEKA, the software used for model building throughout this work: reduced-error pruning, subtree replacement and subtree raising (Kotsiantis, 2013, James et al., 2013). However, as no pruning technique outperforms the other in all datasets (Kotsiantis, 2013), the pruning should be optimized for each different decision tree built (which has been done throughout this work, as explained in the Methods section 3.4.1). In classification each leaf node with be assigned the majority class of its instances, whereas in regression it will be assigned the average of the response of its instances.

One of the most widely used and successful decision tree algorithms is C4.5., proposed by Quinlan (Quinlan, 1993) as it shows a good balance of speed and predictive power (Kotsiantis, 2013, Witten et al., 2011). A more recent version, called C5.0 (no reference available, developed by Ross Quinlan and available in www.rulequest.com/see5-info.html), was developed to improve efficiency in memory usage and computational speed however there are instances where comparative studies show that C4.5 still outperforms, or produces equivalent performance to C5.0 (Rokach and Maimon, 2015, Galathiya, 2012). One of the advantages in C5.0 is that is typically produces smaller rulesets (Galathiya, 2012), hence making it easier to interpret. Decision trees have a very intuitive structure, which makes it easy for a user to understand the decision path that produced each prediction (Freitas, 2013). Additionally, this machine learning method can accommodate high-dimensional data while maintaining the ability to identify and ignore irrelevant variables, which makes this one of the best options for interpretable models for data containing various distinct underlying mechanisms that produce the modelled response. However, decision trees typically obtain relatively low predictive performance (Yousefinejad and Hemmateenejad, 2015, Kotsiantis, 2013) when compared, for example, with SVMs or neural networks (Hastie et al., 2009). Lastly, while non-linearity affects their predictive performance, decision trees appear to be robust to heteroscedasticity and multicollinearity (Kotsiantis, 2013).

In this thesis decisions trees were applied (as a standalone method) exclusively to output data of categorical nature, using the C4.5 algorithm implemented by Quinlan (Quinlan, 1993), so this shall be described in more detail.

Starting from a set of features that describe an output class, the problem originally has maximum entropy (worst separation of classes). The algorithm evaluates the information

gain (or entropy loss) achieved for splitting the data using each available feature, and selected the feature with larger information gain in a greedy fashion. Information gain can be described as the increased ability to separate classes from the previously observed class ratio, as can be calculated with Eq. 2.4.

$$\mathrm{IG, i} = \mathrm{Entropy_{i-1}} - \mathrm{Entropy_{i}} \qquad \text{(Eq. 2.4)}$$

$$\mathrm{Entropy} = \sum_{c=1}^{m} -\frac{N_c}{N_{total}} \times \mathrm{LOG_2}\left(\frac{N_c}{N_{total}}\right) \qquad \text{(Eq. 2.5)}$$

This feature will compose the first node, which will split the data into two subgroups. From this point onwards, C4.5 applies the procedure to all features (including the feature used in the first node) to generate subsequent child nodes that continue to split the subsets originated under each node, so as to maximize class separation. As mentioned earlier for decision trees in general, the extent to which this recursive data splitting occurs is controlled by a stopping parameter (e.g. minimum instances per child node) or by post-pruning (decision tree "simplification"). (Quinlan, 1993, Witten et al., 2011)

2.5.2. Random Forests

The reasoning behind random forests (and any other ensemble method) draws from the empirical value of committee decision making, where predictions are overall more robust.

In random forests, a set of random decision trees are built where the individual trees make individual decisions (predictions) for each instance, and instances are classified according to the majority vote (in classification) or according to an average of the numerical predictions of the members of the ensemble (in regression) (Hastie et al., 2009, Aggarwal, 2014). In the Random Forest algorithm built by Breiman in 2001, in order to build a random forest, bootstrap subsets of training instances are used to train each tree, and a small set of features are randomly selected and made available to the creation of each node. The feature assigned to create the split for each node corresponds to the feature that produces the highest information gain (Equation 2.4), also calculated from entropy (Equation 2.5).

To keep the process unbiased, no pruning is applied to each tree (Breiman, 2001). Producing a group of trees in such manner allows compensating for the instability of single decision trees caused by small changes in the training set (Strobl et al., 2007). As a result, this algorithm works towards reducing the variance in the prediction of the response variable (as many other algorithms do), as well as reducing the bias in that prediction (Breiman,

2001). It assures convergence (by the Strong Law of Large Numbers (Hazewinkel, 1993), whereby, in a sequence of random variables, their averages tend to constant values with probability of 1) and, consequently, it is relatively insensitive to overfitting (Breiman, 2001). Additionally, Random Forests cope with the "small n, large p" issue (small number of instances, large number of features) and, as with decision trees, they can handle scenarios of highly correlated descriptors and can capture non-linear feature-response relationships (Boulesteix et al., 2012, Hastie et al., 2009). However it should be noted that, despite being relatively robust, this algorithm becomes at increased risk of overfitting when there are a small number of informative features among a large total number of variables, and a small feature sampling size is set during training (Hastie et al., 2009). Additionally, negligible gains come from imposing a limit to tree growth (Hastie et al., 2009), which means this should generally be avoided as such constraints are also introducing additional bias. Random Forests are one of the most successful (Biau, 2012) and most widely used algorithms in bioinformatics and chemoinformatics modelling, and a diverse list of examples is provided in the literature (Boulesteix et al., 2012).

As both regression and classification versions of random forest are employed in this work, it is useful to provide additional detail regarding the differences between their implementation, with particular emphasis on how such differences take shape in WEKA. The overall process explained earlier is applicable to both cases, however two main differences exist between them: how features are selected and how the predictions are computed. In classification, features are selected as explained in the previous section on decision trees (2.5.1), where at each point a feature that produces the largest information gain compared to the previous state of the data is selected to form a new decision node. However, while here the corresponding entropy is calculated from class separations achieved with a given tentative split, this cannot be applied to a regression problem where the output is of continuous nature. As a result, WEKA compute information gain from variance measured before and after a given decision split, and the descriptor producing the largest variance decrease will generate the largest information gain and will, consequently, be selected for the current decision node (Information obtained from inspecting the source code for the RandomTree function in WEKA). Regarding how prediction are computed in both modelling scenario, in classification the final predictions are obtained by gathering the majority class votes across the forest, while in regression the final predictions result from taking the average value obtained from the training instances that are allocate to a given final leaf node, and performing an instance-wise averaging these individual predictions across the tree.

2.5.3. Boosted trees

Boosted trees rely on the notion that it is easier to average many rough rules than it is to find a single highly predictive rule, so this technique produces many poorly performing models and combines their output to yield a more powerful committee. As the name implies, this algorithm learns through a technique called boosting where fitting is done by iteratively increasing the emphasis placed on poorly predicted instances. Boosted trees are applicable to both classification and regression problems (Elith et al., 2008, Hastie et al., 2009).

In practice, boosted trees are trained through the sequential fitting of weak decisions trees where, at each step (or boosting iteration), weights are increased in training samples with a poorly predicted response, and decreased for correctly predicted instances. Such tracking of fitting quality is done through a loss function which can adopt different types (e.g. exponential or binomial deviance for classification problems, and absolute loss or squared-error loss for regression, among others). This will force the algorithms to focus on the most challenging examples, as more iterations are performed. After the last iteration, the full set of predictions for each instance is submitted to weighted majority voting to output the final predictions across the dataset. This means that the final predictions will be mainly influenced by the more accurate classifiers in the committee (Hastie et al., 2009, Aggarwal, 2014).

The main attractive feature in boosted trees is that it is able to minimize error rate during learning even for committees formed of near-random classifiers such as decision stumps. Furthermore, compared to simpler approaches such as decision trees, boosted trees provide (often dramatically) improved accuracy, however interpretability and computational cost are sacrificed to achieve that (Hastie et al., 2009).

In a regression context, the boosted trees will be trained by initializing a fitted function with a zero value and assigning the residuals to the observed response Y. Training will then be an iterative process of fitting each subsequent tree not to the actual response Y, but to the current residuals (as a function of the independent features). The new resulting regression tree will then be added to the fitted function, so that the residuals are updated. As in classification, a shrinkage parameter, used to control the boosting learning rate, is applied in association to the output of each iteration (James et al., 2013).

2.6. Multi-label Classification

Up until this point all explanations regarding classification and the corresponding machine learning algorithms were applicable to a training scenario where there is a single response variable (regardless of the number of classes contained in it). This is the traditional machine learning paradigm, where among n training instances (compounds) in the dataset, each instance is assumed to be associated with a single response (called class label). As a result, this is called a single-label problem, and each compound is classed under one label (response), e.g. active or inactive. However, there are cases where instances, due to their complexity, might have various simultaneous responses, which is the same as saying that an instance is associated to a set of various labels rather than just one. By contrast to the previous situation, this is called a multi-label problem. Modelling such complex sets of endpoints will produce a multi-label classifier (Zhang and Zhou, 2014, Carvalho and Freitas, 2009). Formally these two scenarios can be defined as follows:

In a single-label problem, each object is represented by an instance and each instance is in turn associated to a single label. As a result the problem can be formally characterized by an instance space $X = (x_i : 1 \leq i \leq n)$, where n is the number of instances, and a label space $L = (y_j : 1 \leq j \leq q)$, where q is the number of labels, where each instance $x_i \in X$ is associated to a label $y_i \in L$. Thus, each instance represents a property (or set of properties) of an object and each label represents its semantics (Zhang and Zhou, 2014). In traditional supervised learning each instance is assumed to be associated with a single semantic meaning, which means that the main goal here is to learn from a training set $((x_i, y_i) | 1 \leq i \leq n)$ and produce a function $f : X \to L$. (Zhang and Zhou, 2014)

Alternatively, in a multi-label problem, objects have various simultaneous meanings, which means that an instance is associated with a set of labels rather than just one. So, in this scenario the machine learning algorithm is used with the goal of learning a function $h : \mathrm{X} \to (0,1)^{|L|}$ from a training set $((x_i, Y_i) | 1 \leq i \leq n)$ formed by instances assigned with a subset of labels $Y_i \subseteq L$. Hence, for every new instance the trained classifier h(·) will predict $\mathrm{h(x)} \subseteq L$ (Zhang and Zhou, 2014, Luaces et al., 2012).

The interest in learning from multi-label datasets has recently started broadening into a variety of fields, and applications that span from genomics to music. Hence, this research field is quickly evolving (Tsoumakas et al., 2010, Zhang and Zhou, 2014).

Supervised learning from multi-label data can be divided into two main types of task: multi-label classification, which concerns models that produce a bipartition of the labelset into

irrelevant and relevant; and label ranking, in which the model outputs ranked class labels according to the relevance to a given instance (Tsoumakas et al., 2010).

A common methodological approach to handle multi-label classification problems is problem transformation, where an initially multi-label problem is transformed into one or several single-label problems. A regular single-label classifier is then applied to each single-label problem, and the separate predictions from all the single-label classification tasks are finally gathered in the multi-label prediction phase. There are a number of different families of problem transformation methods (Read et al., 2009).

The task of machine learning applied to multi-label problems has two main problems. Firstly, they have an increased computational complexity compared to the single-label counterpart. As consequence, depending on the number of labels, more complex machine learning algorithms may not be practical to use, which obviously hinders scalability. Secondly, the very nature of multi-label data encompasses new levels of uncertainty due to the fact that each instance is associated with an indeterminate number of labels and there can be interdependency between labels (Luaces et al., 2012). Even if label dependency is not known a priori, one cannot safely exclude that possibility, which is why one of the main goals in multi-label classification is to enable the detection of these relationships. Correlations between labels potentially hold important information about the modelled problem, and accounting for this is crucial in facilitating the machine learning algorithm in learning the various responses (Gibaja and Ventura, 2014). As a result, a major goal in multi-label classification is to enable the detection of these relationships.

2.6.1. Binary Relevance (BR)

One of the most widely used problem transformation methods is Binary Relevance (BR), which, as mentioned in the general definition above, decomposes the multi-label problem into a binary problem for each label separately, following a so called one-against-all approach. When applied, the classifier will predict the 0/1 distribution in every separate label, ignoring the information from all the remaining labels (Read et al., 2009, Luaces et al., 2012). The separate predictions from all the single-label classification tasks are finally gathered into one multi-label prediction (Luaces et al., 2012, Read et al., 2009). Formally, BR can be defined by training $|L|$ binary classifiers C_i, $1 \leq i \leq |L|$. Each classifier will be associated to a label l_i and will be used to predict the class of each instance under that label. In practice this means that the classification algorithm uses all the occurring instances

under a given label as positive and all the remaining instances as negative instances (Madjarov et al., 2012).

Even though it is popular, BR has the major drawback of assuming label independence. In practice, by separating the labels one is, in fact, losing potential information; and it has been pointed out that, as a result, predictions produced from BR are likely to contain too few or too many labels, or even impossibly coexisting labels in practice (Read et al., 2009, Luaces et al., 2012). Another problem of using this kind of multi-label method is that one is tacitly assuming that missing labels and negative label observations merge, both of them belonging to the non-positive observations group, so to speak. In other words, one would be considering missing label and negative label observations as the same. This is obviously a flawed assumption. However some authors still argue that despite its obvious flaws, BR-based methods also have valuable features (Read et al., 2009, Luaces et al., 2012), in particular its simplicity and relatively good computational efficiency.

It is widely accepted by the machine learning and data mining community that accounting for inter-label dependency is of paramount importance in several multi-label classification problems. One of the most commonly used and straightforward methods to allow this is the label powerset (LP) method, in which the original multi-label problem is transformed into a single label problem in which each class is a combination of labels (or label subsets) occurring in the dataset. However, since all possible combinations of labels observed in the dataset have to be covered, this usually translates into a large number of different classes. In fact, in most cases the number of possible label combinations (or classes) grows exponentially as dataset size increases, potentially reaching tens of thousands of classes. This is associated with a high computational complexity that is upper-bounded by $\min(|D|,2^{|L|})$, where D represents the number of data points. In practice, this means that this method is suitable only for datasets with a relatively small number of labels (Read et al., 2009).

Turning back to the main drawback of BR, a practical and easily implemented alternative to overcome the assumption of label independence is the Classifier Chain (CC) method, which is able to cope with label dependency (Read et al., 2009). In this technique, the different labels originating from single-label models communicate the learned information to each other, in a sequential fashion. This will be further explored below.

Regardless of any weaknesses, BR still remains in use as the main baseline for multi-label classification. Furthermore, Luaces and colleagues (Luaces et al., 2012) argue that it should not be regarded merely as a baseline for the multi-label classification task, as they demonstrate with synthetic, noisy datasets that for high numbers of labels BR seems to

outperform a CC ensemble. Interestingly, they demonstrated also that with increasing label dependency (even at higher degrees than the current benchmark datasets) the performance of BR is comparable to that of a CC ensemble, which actually goes against what is generally established in multi-label machine learning.

2.6.2. Classifier Chain (CC)

Generally speaking, BR is more known by its shortcomings, and rarely regarded for its advantages. However, this is a very intuitive, simple method associated with comparatively low computational costs, scaling linearly with the increase in the number of labels. Moreover, it is able to optimize various loss functions, training can be performed with any binary learning algorithm, and it can be easily parallelized (Read et al., 2009, Luaces et al., 2012). As already mentioned, on the other side of the spectrum lie methods like LP, which have the advantage of coping with label interaction and transforming the data without any loss of information, but have the disadvantage of being prohibitively expensive during training in many cases. Therefore, it seems logical to use a method that compromises between BR and LP, which is the case of CC. Indeed, as experimentally shown by Read et al. (Read et al., 2009) for 6 benchmark datasets, CC shows only a slight, negligible increase in computational cost while significantly improving the predictive performance in almost all datasets. This was also demonstrated by Luaces et al. (Luaces et al., 2012), who extended the analysis to synthetic datasets distributed along a wider range of numbers of labels. As a result, CC appears to fall in an optimal region in terms of the cost-to-predictiveness trade-off.

As in BR, CC uses $|L|$ binary classifiers, where each classifier deals exclusively with the binary relevance problem for each label. The main difference is that in CC the classifiers are linked in a chain, which extends the feature space of each BR problem (or link of the chain, in this case) by taking into account the 0/1 predicted values of all previous labels. This results in the production of a chain of binary classifiers $C_i, \cdots, C_{|L|}$. The multi-label classification is achieved by |L| steps: at the first link in the chain, C_1 predicts the value (1 or 0, i.e. presence or absence of the label) for the respective label 1, given the training instances available for the classifier associated with that label. This involves computing the first class label's probability: $\Pr(l_1 | x_{1,j})$. For the following links, the prediction is computed for each successive label, given the current classifier's training instances and the predicted label value for the previous labels in the chain. This involves computing the i-th class label's probability: $Pr(l_i | x_{i,j}, l_1, \cdots, l_{i-1})$. By propagating predicted label values from one classifier to the following ones, CC accounts for label correlations (Read et al., 2009); however this

is done without overwhelming the classification algorithms with all possible combinations in the label space, which makes the multi-label classification task much simpler. The main shortcoming of this method is the fact that it relies on the order of the chain. Due to the sequential nature of a CC model, error is propagated through the chain as prediction are fed forward and used as features in the following single-label elements of the CC. As a result, if a poorly learned variable is put in first place it will produce a set of poor predictions which will form a bad source of information for the following variable that are subsequently modelled. On the other hand, this scenario will be drastically different if this initial variable is put in last place, instead, and has a change of receiving additional information produced from other variables which might mean it will be modelled more accurately. The second scenario will more likely produce an overall better performing CC model. As a result of this, to address the effect of label order, if the set of labels is sufficiently small, an exhaustive comparison of all label combinations can be performed to find the optimal ordering. Alternatively, for larger feature sets, this can be overcome by employing an ensemble of CCs (ECC) (Luaces et al., 2012). However, Madjarov et al. (Madjarov et al., 2012) showed that, when compared to each other in terms of the average performance across all available benchmark datasets, ECC does not differ statistically from CC in most of the commonly used evaluation measures, and in some cases it is even outperformed by CC (even though not significantly). Genetic algorithms have also been applied to efficiently search for an optimal arrangement of label. The authors offer a possible explanation for this phenomenon: CC is a stable method, and so ECC does not offer any significant improvement in predictive performance. This, associated with the fact that the training time with CC is much closer to that for BR than ECC, makes CC a very appealing method.

2.7. QSAR Models' Predictive Performance and Reliability

The evaluation of predictive performance is done based on two components – internal and external validation – which both occur after a model has been built. In internal validation, cross validation (CV) strategies are used. While fitting can be improved by merely adding more features, cross validation fitting performance tends to decrease in such situations. The CV strategies can be leave-one-out or leave-many-out, however it has been established that leave-one-out CV is no longer appropriate in most cases and it should, whenever possible, be replaced by leave-many-out CV (also known as k-fold CV), as this gives a more reliable estimation of model generalizability (Yousefinejad and Hemmateenejad, 2015, James et al., 2013). This is due to the trade-off between bias and variance. While leave-one-out CV is less biased than k-fold CV, it has much higher variance. This results from the

fact that the leave-one-out predictions are output by models trained on almost identical training data, so they are highly correlated, which is associated with higher variance. This variance will expectedly be overestimated in relation to the error in the external data, hence not being an appropriate surrogate to be used in model optimization or estimation of generalizability (James et al., 2013).

In cross validation, a user-set number of partitions is applied to the dataset, and each partition is iteratively set aside for testing, while the remaining partitions are used to build a model. The model is then tested on the excluded (testing) partition. This is done in such a way that it allows setting aside every partition exactly once (Alexander et al., 2015). Even though it is useful, internal validation is not sufficient to evaluate predictive performance as it is known to show overly optimistic performance (Alexander et al., 2015, James et al., 2013). Testing the model on external data which was not used in any part of model training and optimization is a more stringent way to estimate the model's true performance, and it is considered the gold standard among the community (Yousefinejad and Hemmateenejad, 2015, Alexander et al., 2015). It is recommended that 20-30% of compounds in the full dataset are set aside for external testing (Yousefinejad and Hemmateenejad, 2015).

2.7.1. Applicability Domain

As pointed out by Eriksson (Eriksson et al., 2003), end users of a QSAR model will only trust its predictions if they are supported by evidence that the chemical space used for training covers newly tested compounds. As a result, any chemistry-response relationship model needs to demonstrate not only good accuracy but also reliability for external predictions. While the former is straightforwardly assessed using performance measures, the latter is addressed by characterizing the model's applicability domain (AD), which consists of the input space (chemical space, in the context of cheminformatics) circumscribed by boundaries inside which the model has reliable and defined performance (Toplak et al., 2014). This has the important role of reducing the propensity for non-apparent extrapolations, i.e. predictions done within input range, however associated with a poorly modelled inner region of such range.

The currently adopted European Chemical Regulation (REACH) promotes the use of in silico models as a replacement for in vitro and in vivo testing for the evaluation of chemical entities. However, results derived from approaches such as QSAR should only be used if the respective model is compliant with four main criteria, among which is the demonstration that the applicability domain of the model appropriately covers new predictions (Madan et

al., 2013). This demonstrates the importance of this component of computational modelling and prompted the exploration of this topic in the context of this thesis' data.

There are several reviews and comparative studies on AD methods available in the literature (Jaworska et al., 2005, Netzeva et al., 2005, Sahlin et al., 2014, Dragos et al., 2009, Sushko et al., 2010a, Sahigara et al., 2012, Mathea et al., 2016), which focus on either distinguishing inliers from outliers, or high accuracy compounds from low accuracy compounds. Contrarily to the modelling task where a response variable can be used to assess the predictive ability of the model, there is no response variable for the true inclusion in the AD, given its subjective nature. Therefore, the characterization of a model's AD is exploratory by nature. While there is no way of objectively determining the accuracy of forecasts on inclusion/exclusion criteria of new queries within the AD, one is able to estimate the utility of a certain AD in a real-world scenario by applying it to naïve data.

Much like machine learning methods, available AD characterization methods are broadly divided into two types: unsupervised AD, which relies solely on descriptor space and supervised AD, which uses the relationship between descriptors and the output variable. Examples of unsupervised AD are range-based methods (coverage determined by the range of each individual descriptor), convex hull (coverage determined by the smallest convex area that contains all training compounds), distance-based methods (coverage determined by distance to training set) and density-based methods (coverage defined by density of training space in areas affecting new compounds). Unsupervised methods have many limitations that stem from the difficulty to capture the multidimensional structure of the data. Multiple problems arise, such as the distortion of distance measures by the presence of correlation between descriptors, or by the existence of a highly dimensional space. On the other hand, if descriptors are used individually a compound might be deemed well covered by some descriptors and not by others. In addition they rely on the assumption of a smooth input-output landscape where predictions are less reliable as compounds distance themselves from training space, in a somewhat proportional manner. This is easily violated in a real world scenario, where small input changes lead to a dramatic change in the output response, and where roughed patches on the structure-activity landscape are more difficult to properly capture by a machine learning algorithm. For example, a compound might be well covered by descriptor space but might belong to a completely different class, and a (perceivably) remote compound is not necessarily unreliably predicted.(Mathea et al., 2016, Netzeva et al., 2005, Jaworska et al., 2005, Kaneko and Funatsu, 2017)

As a result, supervised AD is more capable of detecting such situations as it uses the machine learning task to infer on the reliability of a new compound. As a consequence this

is generally regarded as being superior to unsupervised approaches, as it tries to capture directly the propensity for unreliable predictions. In this context reliability can be derived from the distance to decision boundaries (for classification), the level of agreement in an ensemble of models or the prediction probability. (Mathea et al., 2016)

2.8. QSAR Models for the Prediction of the Drug Transport

This section will provide an overview on the different approaches used, to date, to models transporter data. The same will be done for Volume of Distribution in Chapters 6 and 7, as in these chapters a direct comparison to the literature is beneficial during the discussion of results obtained.

Sedykh et al. (Sedykh et al., 2013) reported the QSAR modelling of various ATP-Binding Cassette transporters (ABCs) (all 4 transporters considered in this thesis) and Solute Carriers (SLCs) (PEPT1, OCT1 and OATP2B1) , and by accounting for the models' applicability domain they were able to generate overall high predictive performance across all transporters. Majority class undersampling was applied to PEPT1 and MRP2 data, where only the least similar compounds to the minority class were removed.

Very recently (during the same period the work in this thesis was being developed), Ose et al. (Ose et al., 2016) reported Support Vector Machine (SVM) classification QSAR models for ABC and SLC data, however these were built on from data merged from physiologically related transporters (OATP1B1+OATP1B3; OCT1/2+MATE1/2-K; MRP2/3/4) into a single response class. They trained an SVM classification model on a very limited number of descriptors (ranging from 4 to 7) which were selected partly selected by empirical criteria and partly by feature selection). These models showed high prediction performance, however they are not applicable to predict whether a compound is likely to be a substrate of any one particular transporter. Additionally they have modelled BRCP1 and P-gp, individually, where the models were able to reach good predictive performance for the latter, but low precision for the former.

After the development of the work in this thesis, Shaikh (Shaikh et al., 2017) also reported on modelling of a variety of ABC and SLC transporters and created ensemble QSAR models for each transporter. Data was gathered from the same source as used in this thesis (i.e. Metrabase (Mak et al., 2015)) and PEPT1, OCT1, OATP1A2, OATP1B1, OATP1B3, OATP2B1, BCRP, P-gp MRP1 and MRP2 substrate and non-substrate datasets were modelled. However note that these datasets were completed with putative non-substrates

(to overcome imbalance) that corresponded to compounds that undergo passive permeation. Putative non-substrates corresponded to 0-42% of the total dataset, which means that they are more frequent than actual non-substrates in some cases. They have also applied some physicochemical filters to make the chemical space more tractable to the training step, and have used protein-derived descriptors that encode both intrinsic transporter properties and protein-ligand interaction. They report the performance of the top models, however this appears to have resulted from a selection based on the external validation dataset. This incurs in the risk for overfitting and performance values obtained by models tested in the same dataset that was used previously for the selection of such models should be analysed conservatively.

It should be noted that using passive permeation as counter-examples trained against substrates of any given transporter has an associated risk of producing a model that merely learns to distinguish passive diffusion from any kind of active transport (being insensitive to the actual transporter being modelled).

2.8.1. The ABC Superfamily

P-glycoprotein (P-gp)

Various QSAR models of P-gp substrate prediction have been carried out and are generally associated with higher predictive power than structure-based methods of pharmacophore (effluxophore) models (Broccatelli, 2012).

Initially, a relatively small amount of data was used (N=195). Lima et al (de Cerqueira Lima et al., 2006) used this dataset annotated with a range of molecular descriptors to produce a range of QSAR models and obtained an accuracy in external data (ca. 25% of the data) of 81% for their best model (which was trained using a support vector machine (SVM)) (de Cerqueira Lima et al., 2006). Cabrera and colleagues added observations to this dataset and applied additional quality filters where borderline substrates and bound but non-transported compounds were removed. They reported 77.5% accuracy in external compounds from their linear discriminant analysis model. Huang et al (Huang et al., 2007) and Wang et al (Wang et al., 2011) also modelled the dataset from Cabrera et al. and, not surprisingly, both produced SVM models of much larger accuracy in external data (90% and 88%, respectively) set aside from the full dataset.

Crivori et al (Crivori et al., 2006) trained a partial least squares discriminant analysis model on a relatively small set of data (N=53) and upon testing it in a significantly larger external dataset (N=272), they were able to report 72.4% accuracy. This is perhaps partly caused

by the fact that they have assigned to the substrate group two compounds that would be considered “indeterminate” cases, as they had an efflux ratio (ER) < 2.

Gombar et al (Gombar et al., 2004) also trained a discriminant analysis model but on a larger proprietary dataset (still relatively small, N= 95), having applied more stringent parameters than Crivori et al. when assigning data to either substrates or non-substrates (substrates if ER > 2, and non-substrates if ER <1.5). In this case, despite having used more data for the training than Crivori et al., which would theoretically lead to better interpolation, their model shows a worse specificity than a model trained with half of the data (73.9% versus 75.0%, respectively). Even though overall accuracy is deceivingly larger in Gombar et al. (86.2%), this is actually a worse model as it shows larger imbalance between sensitivity and specificity – for a precise definition of these terms see Chapter 3. The main reason here appears to be related to class representation and class imbalance, where Crivori et al. used 58.5% non-substrates, whereas Gombar et al. used only 32.6% for training. This demonstrates the relevance of class imbalance towards predictive power, which can counteract applied criteria to improve data quality.

Gupta et al. (Gupta et al., 2010) used a very large proprietary training set of 24,995 compounds. These authors applied yet another threshold criterion for the classification of compounds where compounds are considered substrates if ER ≥ 2.5. The allocation into training and test sets was done in such a way that assures maximum variety among the same group of compounds. In this work, a number of different machine learning algorithms were used to train the models (SVM, recursive partition forest, C5.0 and random forest). An additional 18,413 new compounds were used for testing, and this yielded somewhat varied performances across the different machine learning algorithms. SVM was surprisingly the lowest performing model, with very low specificity (52%) and C5.0 was the best performing model with high sensitivity and specificity (85% and 77%, respectively).

Desai et al (Desai et al., 2013) reported different chronological variations of a bagging trees P-gp efflux model developed on a large proprietary dataset, in which they allocated compounds into substrates and non-substrates using yet another ER threshold value (i.e. substrate if ER > 3, otherwise non-substrate). Different time cohorts of data (ranging from 863 to 1945 compounds) were used to build the bagging model and they showed overall 80% predictive accuracy. Their QSAR model (which was built from manual in vitro P-gp efflux assay) proved to be more reliable as a predictor than an alternative automated P-gp efflux assay, and thus the latter was discontinued and replaced by the QSAR model.

Broccatelli (Broccatelli, 2012) took a different choice from other works relying on open-access data, and tested P-gp efflux modelling with a single, high quality source. This tested

the feasibility of prioritizing the minimization of noise at the cost of loss of information. He built several models using a small dataset of 150 compounds whose efflux measurements were available from the Netherlands Cancer Institute, and were exclusively determined on MDCK-MDR1 Borst cell lines. His best model (which was built using a random forest algorithm) showed 84% accuracy and a high balance between sensitivity and specificity (80% and 86%, respectively) in an external dataset of 37 compounds. This model showed the highest predictive power, when compared to other works that also used highly consistent datasets derived from one single source of efflux ratio data, namely Desai et al.'s and Gupta et al.'s work; even though these used a training set one to two orders of magnitude larger. This might appear counterintuitive, however this might be due to the fact that these two other works used proprietary data, which may mean that they cover a larger amplitude of chemical space that originates from the normal practice of companies to actively explore new scaffolds and therapeutic families. This might hinder the prediction performance.

Breast Cancer Resistance Protein (BCRP)

Contrarily to P-gp, BCRP has much fewer QSAR models reported in the literature, and all published in recent years. Hazai et al.(Hazai et al., 2013) and Gantner et al.(Gantner et al., 2013) built BCRP efflux QSAR models (using, respectively, SVM and MLR/ LDA methods) on datasets of similar size (263 and 262 compounds, respectively). They have obtained comparable performance (73% and 74.5%, respectively), even though the former allocated considerably more data to training (N=223) than the latter (N=164). Zhong et al. (Zhong et al., 2011) used a smaller set (N = 137) of compounds to also train an SVM model, which yielded an even better accuracy of 85% in 40 external compounds, which is probably due to the fact that training was paired with a round of GA feature selection and a parameter optimization method (conjugate gradient).

More recently, Ose et al. (Ose et al., 2016) built a multi-label binary relevance model for a number of SLC and ABC transporter endpoints. For each transporter they have trained a separate SVM model where experimentally confirmed substrates are classified against expert-suggested non-substrates. The model predicting BCRP was able to identify 21 out of the 27 available substrates, however looking at the full set of predictions shows that the model predicted a total of 56 compounds as being substrates. As the 35 exceeding compounds may or may not be substrates, it is unclear whether the model has properly learned the BCRP endpoint or it just overfitted (as it classified more than half of the external dataset as substrates).

Multidrug Resistance Proteins 1 and 2 (MRP1 and MRP2)

There are very few QSAR models on either MRP1 or MRP2 substrate data. Initially a model of MRP2 built on 1204 putative substrates and non-substrates was reported, indirectly derived from the correlation between cytotoxicity and MRP2 mRNA levels. This however did not yield good results in 44 external, experimental observations which showed a sensitivity well below 50% (Pinto et al., 2012).

Additional attempts to model MRPs, reported by Ose et al. (Ose et al., 2016), Sedykh et al. (Sedykh et al., 2013) and Shaikh et al. (Shaikh et al., 2017) have been discussed at the beginning of section 2.8, as these work report on multiple transporters simultaneously.

2.8.2. The Solute Carriers (SLCs) Superfamily

As SLCs have very few ligand-based QSAR models reported in the literature, these have been discussed in the same section, at the beginning of section 2.8, because applicable works typically address different transporters at the same time. These are namely the works by Ose et al. (Ose et al., 2016), Sedykh et al. (Sedykh et al., 2013) and Shaikh et al. (Shaikh et al., 2017).

2.9. Summary

Considering this thesis is in the intersection of two disciplines - pharmacokinetics and machine learning - this chapter provided an overview of the theoretical principles behind the machine learning methodology (QSAR modelling) applied to address different pharmacokinetics questions posed throughout this work.

In order to develop a QSAR model, a series of steps should be followed, comprising data curation, calculation of chemical features, feature selection which finally lead to model development using statistical machine learning. According to the nature of the output variable (continuous versus categorical) different feature selection and machine learning algorithms can be employed. Regarding the latter, one of the main points of focus of this thesis is multi-label methods, which allow harnessing links or correlations between different distribution phenomena, ultimately aiding in the ability to model such phenomena.

After a model has been developed, its ability to derive useful predictions should be assessed, and the different types of performance assessment were discussed. Lastly, beyond determining the level of accuracy in predictions, it is important to characterize the model's applicability limits, defined as the applicability domain.

Lastly, a summary of the state of the art in terms of QSAR models applied to the various pharmacokinetics endpoints addressed by this thesis was provided.

3. Methodology and Workflow

3.1. Datasets

3.1.1. ABC Efflux Dataset

A dataset of 1493 compounds was compiled from the substrate data available in the Metrabase database (Mak et al., 2015) (accessed in October 2014) for six ABC transporters: BCRP1, MDR1, MRP1, MRP2, MRP3 and MRP4. All instances were divided into two classes: substrates and non-substrates. The collection of SMILES provided was checked for repetitions and isomers using ACD Labs, and mixtures were removed. Repetitions were merged and, for cases of conflicting information, the principle of minimum evidence was applied, by which all compounds with at least one case of reported substrate property were regarded as potential substrates and so, they were classified as substrates. This is a valid approach considering that all the initial data collected from Metrabase was selected based on quality standards (Mak et al., 2015).

The resulting dataset contained 1493 compounds which showed a negligible imbalance in class label distribution for larger transporter classes, i.e. BCRP1, MDR1, MRP1 and MRP2 compounds, with the substrate to non-substrate ratio of 1.7, 1.3, 1.0 and 1.2, respectively. However, for the smaller transporter classes, namely MRP3 and MRP4, the ratio was around 2.5, which led to insufficient number of non-substrates for modelling and validation. Therefore, these two transporters were eliminated and the remaining four transporters were investigated, using a final dataset of 1462 compounds spread across transporters as shown in Figure 3.1.

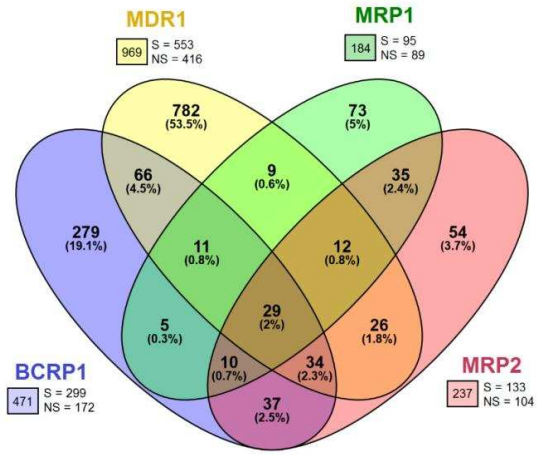


Figure 3.1. Schematic summary of transporter overlap represented in the Venn diagram. Below each transporter label are the total number of instances (in a square) in the full dataset, and the corresponding number of substrates and non-substrates. S: substrates, NS: non-substrates.

3.1.2. SLC Uptake Dataset

Substrate and non-substrate data was retrieved from Metrabase (Mak et al., 2015) (accessed in November 2015), and the binding profiles of all available SLC transporters were collated. This corresponded to OATP1A2, OATP1B1, OATP2B1, OATP1B3, OCT1 and PEPT1. Even though data on the Apical Sodium Dependent Bile Acid Transporter (ASBT) was also available for analysis, this was removed later due to lack of sufficient amount of data to allow acceptable external model testing.

Regarding the annotation of molecular descriptors in this dataset, this was done using the structures provided as SMILES codes were analysed and validated using ACD labs. All duplicates and pairs of isomers were checked using ChemSketch. Duplicated entries were merged when exhibiting agreeing responses, and when contradicting responses were found (4.3% of the observations) the respective entry was annotated as substrate, following the principle that, if a compound has at least one substrate report, it is likely to be a substrate, following the same reasoning applied to the ABC Efflux Dataset (Section 3.1.1). As a result, substrate/non-substrate data for a total of six transporters were used for QSAR modelling, where each transporter corresponds to a class label in the multi-label classification task carried out in this work.

The dataset had a total of 760 unique compounds spread across 980 instance-label cases distributed as shown in Table 3.1. The full compound vs labels matrix was 21.5% filled in and had a label cardinality of 1.3 (i.e., on average, there are 1.3 labels per instance).

Table 3.1. Distribution of substrates (S) and non-substrates (NS) across the different transporters in the SLC dataset.

	S	NS
OATP1A2	55	24
OATP1B1	95	37
OATP1B3	58	26
OATP2B1	47	64
OCT1	159	88
PEPT1	246	81

3.1.3. Volume of Distribution Dataset

The Vd dataset compiled by Obach et al (Obach et al., 2008) was used in the QSAR modelling. This dataset is composed by Vd measurements obtained exclusively from human intravenous administration, in steady state. The SMILES codes for the compounds in the dataset were retrieved from the provided names using the pubchempy module in

python. Retrieved SMILES were checked against a separate retrieval operation of CAS-to-CID search on DrugBank, followed by CID-to-SMILES conversion using the online tool available in PubChem. The few mismatching cases (N=9) found when comparing the two sources of SMILES were clarified through manual retrieval.

The final list of SMILES codes was then standardised using the MolVS python package, which entailed standardisation of chemotypes and tautomers. A comparison of canonical SMILES allowed identifying pairs of repeated 2D structures where pairs of isomers were kept (given that 3D descriptors will be used); otherwise, for pairs of canonical + isomeric structures, only one instance was kept (based on the availability of observations from the included physiological variables in the dataset – see below). The final dataset was composed of 665 compounds.

Regarding the annotation of the dataset with physiological descriptors, given the physiological implication of protein-mediated transport, phospholipidosis (PL), and plasma protein binding (PPB) in the distribution of numerous compounds, these were added as physiological descriptors (PDs), used alongside a set of molecular descriptors (MDs). Specifically, twelve PD variables were added: drug-induced PL, ABC transport (mediated by P-gp, BCRP, MRP1 and MRP2), SLC transport (mediated by PEPT1, OCT1, OATP1B1, OATP1B3, OATP1A2 and OATP2B1) and PPB.

Transporter Binding Data was retrieved from the ABC and SLC datasets described in 3.1.1 and 3.1.2, and compounds in the Vd dataset were annotated with a binary response. The same was done for Drug-Induced PL Data which was retrieved from different sources available in the literature(Lowe et al., 2012, Goracci et al., 2013, Orogo et al., 2012, Bauch et al., 2015, Muehlbacher et al., 2012) as well as from ChEMBL (removing repeated observations from the previous references). From these, Goracci et al(Goracci et al., 2013) and Lowe et al (Lowe et al., 2012) were regarded as the gold standards, as their data is obtained from electron microscopy measurements (the highest quality source for phospholipidosis data). The remaining sources are herein termed as “secondary”. As a result of this criterion, whenever the full set of measurements for a given compound shows conflicting responses, the responses from those two sources are kept and the remaining conflicting data is ignored. When no information is provided from any of these two sources, information from any of the other sources is accepted, given that two or more competing observations for the same compound have to agree in order to be accepted. For 5 instances there were no agreement between multiple secondary sources, so their PL observations were discarded.

Whenever applicable, phospholipidosis or transporter information associated to a given isomer is assigned exclusively to the corresponding isomer entry in the Vd dataset. Otherwise, it is assigned to the corresponding non-isomeric equivalent entry.

As the number of entries in the Vss dataset where an experimental response for transport or PL could be retrieved were limited, the non-existent responses were completed with predictions. Each of the 10 transporter variables were filled in with the output produced by two multi-label models (chapter 4 and 5) applied in a preprocessing step. This was also applied to PL, where predictions were obtained from a previously trained model on the benchmark PL dataset curated by Goracci. This model was trained prior to the initiating the work in this chapter, by using physicochemical descriptors (obtained as explained in section 3.2) and random forest classification paired with prior greedy search feature selection, implemented in WEKA (Hall et al., 2009) under default conditions. These conditions were selected after preliminary optimization, based on the highest performance in an internal 10-fold cross validation on the training set (not using the testing set). Given that the dataset used is unbalanced, a cost of 2 was assigned to the false negatives (while false positives remained with the default cost of 1) during training. The model was built on 80% of the dataset and tested on the remaining 20%, showing high sensitivity (0.857) and specificity (0.711) on the test set.

To allow differentiation of the transporter predictions and the PL predictions according to their quality, they were used in the form of class probabilities (rather than categorical class predictions). A schematic representation of this process can be found in Figure 3.2. Here, for each transporter, experimental observation is annotated as substrate (1) or non-substrate (2) – shown in black. All missing experimental observations (e.g. compound #1 for BCRP1) are completed with the predictions probabilities drawn from the substrate class of each transporter's classification model (i.e. probabilities up to 0.5 represent a predicted non-substrate, probabilities above 0.5 represent a predicted substrate) – shown in blue.

Compound	BRCP1	OATP1B1
#1	0.2	1
#2	1	0.87
…	…	…
#N	0	0.14

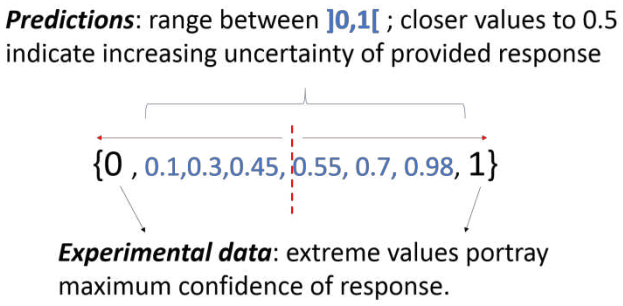


Figure 3.2. Completion of missing transporter binding data with the prediction probabilities obtained from the different respective multi-label classifier.

Lastly, regarding the annotation with Plasma Protein Binding (PPB) Data, this predictor was submitted to the same procedure of completing missing observations with predicted data. For this reason, a prior step of PPB modelling had to be carried out. In order to maximize the achieved predicted power, the PPB data provided by Obach et al was not used and, instead, a larger dataset was used. The PPB dataset deposited in ChEMBL by AstraZeneca was used the single source of data (assay reference CHEMBL3301361) being composed of 1614 compounds. No other data source was added onto this dataset as this was deemed sufficiently large (relative to the scale of the Vss dataset) and doing so reduces the chance of noise that results from inter-laboratory experimentation. The data was modelled using physicochemical descriptors (obtained as explained in section 3.2) and a random forest model (of 200 trees, optimized by 10-fold CV) paired with greedy search pre-processing feature selection. Eighty percent of the data was used for training, and the resulting model yielded a 7.9% mean absolute error in the test set; PPB predictions obtained were used to fill in missing data.

3.2. Molecular Descriptors

All molecular descriptors used as input variables throughout this work were calculated using ACD/labs logD and Molecular Operating Environment (MOE 2013 in chapters 4, 5 and 8; MOE 2015 in chapters 6 and 7). Prior to any calculation, input structures obtained in form of SMILES codes were washed and standardized. As a portion of descriptors calculated in MOE are dependent on the 3D conformation of a compound, all structures were submitted to a minimization protocol beforehand. An initial molecular mechanics minimization was performed (further information on the used method is provided in the Methods section of each experimental chapter), followed by a subsequent refinement with quantum mechanics minimization (using the PM6 method).

No single charge-assignment method was selected over any other, across homologous descriptors, as it has been shown that different charge assignment methods have led to variable success in modelling different datasets in the past (Mittal et al., 2009). This allows a data-driven selection of charge-related molecular descriptors using PEOE vs PM6 methods, as well as various descriptors derived from semi-empirical methods, AM1, PM3 and MNDO.

All invariant or mainly empty descriptors were excluded, as well as repeated and spatial coordinates-dependent descriptors. Descriptors with predictions of activity/response endpoints (such as mutagenicity) were also excluded. pKa and pKb values, calculated for

the most acidic and basic species, were used to calculate the ionized fraction of acid, base and zwitterion at 7.4, as well as the unionized fraction. After this, pKa and pKb were excluded.

3.3. Feature Selection

Five feature selection techniques were employed throughout the work presented in this thesis, and the underlying theory that explains how they function has been given in section 2.3. All feature selection methods were run using the training set only, and they can be divided into three filter methods, namely Genetic Algorithm search (GA), Greedy Stepwise search (GS) and ReliefF (RfF); and two wrapper methods, namely the C4.5 Decision Tree-Genetic Algorithm (C4.5-GA) and Random Forest-Greedy Stepwise search (RF-GS). These were implemented using the popular data mining tool WEKA (version specified in each chapter), following the setting described below.

Filter methods were implemented with the CfsSubsetEval attribute evaluator paired with GA and GS search algorithms. CfsSubsetEval is WEKA's implementation of correlation-based feature selection evaluation, which scores features by rewarding strong correlations to the dependent (class) variable, and penalizing strong correlations to other features (Witten et al., 2011).

For the GA feature selection, the GeneticSearch algorithm was used, being set for 0.8 and 0.01 crossover and mutation probabilities, respectively; and both the population size and the number of generations were set to 100, to allow sufficient exploration of the feature space. GS was carried out using the GreedyStepwise, using the default settings.

RfF was carried out with the ReliefFAttributeEval, which does its own evaluation (i.e. this was not paired with CfsSubsetEval), and was run using default settings which coincide with previously reported used settings (Spolaôr et al., 2013).

The wrapper methods were implemented with ClassifierSubsetEval where the two search algorithms (GS and GA) were combined with two classifiers (respectively RF and C4.5 decision trees, using the RandomForest and J48 implementations in WEKA) to run the two resulting wrapper methods: RF-GS and C4.5-GA. In C4.5-GA, the settings for the wrapper GA were the same as the ones used for the filter GA. As for the C4.5 classifier within the wrapper, the pruning method was optimized by an internal (using the training set) 10-fold cross validation. When applicable, the confidence factor was optimized in a range between 0.1 and 0.5 (with a 0.1 step). All other parameters within C4.5 were set to default values. In the RF-GS method the trees were limited to a maximum depth of 3, as the focus is tree

number not tree depth. The number of trees (ranging from 1 to 25) was optimized using the internal 10-fold cross-validation root-mean squared error.

To minimize local-minima effects that have been particularly reported for GAs (Shahlaei, 2013), for all feature selection methods an internal 10-fold cross validation was re-run multiple times (the exact number will be specified when applicable, in each chapter) using different random seeds.

3.4. Machine Learning Algorithms

Throughout the work presented in this thesis, from chapter 4 to 8, inclusive, the following machine learning methods were employed to model different endpoints, using Weka (Hall et al., 2009) (version 3.6 or 3.8, specified in each chapter).

3.4.1. Decision Trees

Decision trees were built using the C4.5 implementation in Weka – J48. The optimal pruning method was selected from the three available methods (reduced-error pruning, subtree raising and subtree replacement) based on the lowest Mean Absolute Error (MAE) in an internal 10-fold cross validation. When applicable (i.e. when the pruning method was not reduced-error pruning), the confidence factor was optimized in a range between 0.1 and 0.5 (with a 0.1 step), using an internal 10-fold cross validation as well. All other C4.5 parameters were set to default values.

3.4.2. Random Forests (for Classification and Regression)

Random Forest (RF) and Boosted Regression Trees (BT) were used. Both were implemented in WEKA using the RandomForest function and the AdditiveRegression method wrapped around the RandomTree learner. For the tuning of the algorithm's parameters in both cases, the optimal parameter values were selected based on the lowest Mean Absolute Error (MAE) in an internal 10-fold cross validation using the training set. For RF tuning, the number of trees was optimized in a range between 100 and 1000 (considering increments of 100). As for BT, the number of randomly sampled features at each node was set to 9 (the same value as used by default in the RF algorithm), the number of iterations was optimized between 100 and 1000 (again, in increments of 100) and

shrinkage was optimized between 0.05 and 1 (in increments of 0.05). All other parameters in both algorithms were used as default in Weka.

3.4.3. Boosted Trees

Boosted classification trees (BCT) were trained by wrapping the boosting algorithm implementation in WEKA, multiBoostAB, around the C4.5 decision tree algorithm, J48, and were tuned using an internal 10-fold cross validation. The conditions for the embedded J48 trees were inherited from the previously optimized J48 models within the current chapter, the number of committees (or iterations) was optimized ranging from 10 to 100, and the number of subcommittees was set to the squared root of the committee size as recommended by the author(Webb, 2000).

Regarding the boosted regression trees (BRT) the AdditiveRegression wrapped around the RandomTree learner was used. For the tuning of the algorithm parameters, the optimal parameter values were selected based on the lowest Mean Absolute Error (MAE) in an internal 10-fold cross validation using the training set. The number of randomly sampled features at each node was set to 9 (the same value as used by default in the RF algorithm), the number of iterations was optimized between 100 and 1000 (in increments of 100) and shrinkage was optimized between 0.05 and 1 (in increments of 0.05). All other parameters in both algorithms were used as default in Weka.

3.5. Predictive Performance Evaluation Measures

3.5.1. Evaluation Measures for Classification

As established earlier in section 2.6, single-label and multi-label data have different structures and, as a result, the models they generate have to be evaluated using performance measures designed for either a unique endpoint label or multiple coexisting labels, respectively.

Single-label model assessment

The single-label performance measures used for single-label model assessment are defined below (Eriksson et al., 2003), where TP, TN, FP and FN stand for the numbers of true positives, true negatives, false positives and false negatives, respectively (Fawcett,

2006). The measures listed below are called Sensitivity (Sen), specificity (Spe), Matthew's correlation coefficient (MCC), and the geometric mean between Sen and Spe (G-mean).

$$\mathrm{Sen} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \qquad \text{(Eq. 3.1)}$$

$$\mathrm{Spe} = \frac{\mathrm{TN}}{\mathrm{TN+FP}} \qquad \text{(Eq. 3.2)}$$

$$\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP+FN})\times(\mathrm{TN+FP})\times(\mathrm{TP+FP})\times(\mathrm{TN+FN})}} \qquad \text{(Eq. 3.3)}$$

$$\mathrm{G-mean} = \sqrt{\mathrm{SEN}\times\mathrm{SPE}} \qquad \text{(Eq. 3.4)}$$

Multi-label model assessment

Several multi-label predictive accuracy measures were used, namely the harmonic mean between precision and recall (F1), Precision (P) and Recall (R), calculated according to Tsoumakas and Katakis (Tsoumakas and Katakis, 2007, Tsoumakas et al., 2010). Hamming Loss (HL) was used solely to monitor the impact of each label on the multi-label model's predictive performance, during model building.

$$\mathrm{Hamming\ Loss} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\frac{|\mathrm{Y_i}\Delta\mathrm{Z_i}|}{|\mathrm{L}|} \qquad \text{(Eq. 3.5)}$$

$$\mathrm{F1} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\frac{2|\mathrm{Y_i}\cap\mathrm{Z_i}|}{|\mathrm{Z_i}|+|\mathrm{Y_i}|} \qquad \text{(Eq. 3.6)}$$

$$\mathrm{P} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\frac{|\mathrm{Y_i}\cap\mathrm{Z_i}|}{|\mathrm{Z_i}|} \qquad \text{(Eq. 3.7)}$$

$$\mathrm{R} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\frac{|\mathrm{Y_i}\cap\mathrm{Z_i}|}{|\mathrm{Y_i}|} \qquad \text{(Eq. 3.8)}$$

In these measures, Y_i and Z_i correspond to the set of observed and predicted labels, respectively, for the i-th compound, N corresponds to the number of compounds (instances) in the dataset, and |L| corresponds to the number of modelled labels. The Δ symbol denotes the symmetric difference between two sets of label values (observed and predicted, in this case), which is equivalent to the exclusive-or (also known as "XOR") boolean operation.

To overcome bias coming from unbalanced classes, a balanced accuracy (bACC) was used when assessing predictive performance. This measure consists of the average G-mean across every label j (which, in turn, can be considered as the single-label balanced accuracy). To evaluate the predictive performance considering the balance between the two classes across instances, ΔPR measures the average deviation in precision and recall between substrates and non-substrates.

$$\mathrm{bACC} = \frac{1}{|\mathrm{L}|}\sum_{\mathrm{j}=1}^{|\mathrm{L}|}\sqrt{\mathrm{Sen_j} \times \mathrm{Spe_j}} \quad \text{(Eq. 3.9)}$$

$$\Delta\mathrm{PR} = \frac{|\mathrm{P_S} - \mathrm{P_{NS}}| + |\mathrm{R_S} - \mathrm{R_{NS}}|}{2} \quad \text{(Eq. 3.10)}$$

3.5.2. Evaluation Measures for Regression

The measures used for assessing the predictive performance are the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), coefficient of determination (R^2) and the Geometric Mean Fold Error (GMFE), calculated as defined in the literature (Polishchuk et al., 2016, Alexander et al., 2015, Freitas et al., 2015). Additionally, the Mean Fold Error (MFE) and the percentage of data within 2- and 3-fold error (FE) thresholds were calculated for the best final models within each applicable chapter. Consider that predicted, $\hat{y}$, and observed, y, values are log-transformed.

$$\mathrm{MAE} = \frac{\sum|(\mathrm{y} - \hat{\mathrm{y}})|}{\mathrm{N}} \quad \text{(Eq. 3.11)}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum(\mathrm{y} - \hat{\mathrm{y}})^2}{\mathrm{N}-1}} \quad \text{(Eq. 3.12)}$$

$$\mathrm{R}^2 = 1 - \frac{\sum(\mathrm{y} - \hat{\mathrm{y}})^2}{\sum(\mathrm{y} - \bar{\mathrm{y}})^2} \quad \text{(Eq. 3.13)}$$

$$\mathrm{FE} = \mathrm{Antilog}_{10}(|\mathrm{y} - \hat{\mathrm{y}}|) \quad \text{(Eq. 3.14)}$$

$$\mathrm{GMFE} = \mathrm{Antilog}_{10}(\mathrm{MAE}) \quad \text{(Eq. 3.15)}$$

3.6. Applicability Domain (AD) and Activity Cliffs

For any QSAR model, it is necessary to define the domain of applicability to ensure its reliability in the prediction of properties of compounds in an external, independent dataset (from a data source different from the one used to build the model). To determine the AD, the distance to the model based on the standard deviation (STD) of the predicted values (or labels) from an ensemble of various models was used, as this has been shown to be the most successful method in quantifying predictive reliability across chemical space in the data (Sushko et al., 2014, Dragos et al., 2009, Sushko et al., 2010a, Tetko et al., 2008, Tetko et al., 2013). This technique capitalizes on the concept that the disparity between predictions computed from a group of models (ensemble) is a direct consequence of prediction reliability. A small standard deviation will equate to highly reliable predictions, whereas a larger value signals unreliable predictions. It has been demonstrated that the disagreement between models leads to a better separation between reliable and unreliable predictions compared to traditional structure-based measures (Tetko et al., 2013).

An ensemble of models is trained (independently from the actual QSAR models) using random samples of training set data, each sample comprising 80% of the training set compounds.

$$\mathrm{STD} = \sqrt{\frac{\sum(\mathrm{y_m} - \bar{\mathrm{y}})^2}{\mathrm{N}-1}} \quad \text{(Eq. 3.16)}$$

STD values are calculated for each compound using Equation 3.16 above. Here, y_m is the class label prediction using model m and $\bar{y}$ is the average of all prediction outputs for this compound by N models. For classification models (which is the case here) the class label predictions y_m take the form of probabilities. By setting increasingly larger STD thresholds (with increments of 0.05), which can also be perceived as increasing distance to the model's reliability core, more compounds become included in the model's AD. By performing this kind of scanning through the model's space, one is able to establish a profile of reliability (measured in % correct predictions, otherwise called accuracy) as a function of STD.

To further explore a model's domain of applicability, it is useful to identify cases that might appear covered by the training space but are actually associated with a high level of predictive error. A well-known example of this are activity cliffs.

To search for possible activity cliffs, the similarities between all pairs of compounds were calculated using the well-known Tanimoto coefficient (Tc) applied on 1024 bit Morgan circular fingerprints (equivalent to the extended connectivity fingerprints [ECFP], calculated using the RDkit module in python), for a radius of 2. Following the criteria for activity cliffs used by several authors (Iyer et al., 2013, Wassermann et al., 2011, Stumpfe and Bajorath, 2012), activity cliffs corresponded to compound that have a different class than the majority class of the corresponding 3 nearest training neighbors, which must all show a tanimoto coefficient > 0.55 to the analyzed compound. This threshold has been reported as a sensible value above which compounds are visibly similar (Iyer et al., 2013, Stumpfe and Bajorath, 2012, Wassermann et al., 2011).

3.7. Visualization

In order to explore the distribution across chemical space of different datasets (or subportions of it) used throughout this work, t-Distributed Stochastic Neighbor Embedding (t-SNE) (Maaten and Hinton, 2008) was chosen as the multidimensional scaling (MDS) technique. This technique is one of the most successful in conserving the multidimensional structure of the data during its projection into a low-dimensional plot (Maaten and Hinton, 2008). t-SNE starts by taking the all the pairwise Euclidean distances in high-dimensional space and converting them into conditional “proximity” probabilities, which can be seen as similarities. These probabilities (or similarities) are computed such that probability of a given b (or similarity between a and b) corresponds to the probability of b being picked as a neighbor of a, with neighbours being picked under a Gaussian probability density function centered at a. This means that in close high dimensional proximity the conditional probability between two points is high and becomes infinitesimal for distance points. The same is done for low dimensional space, where low-dimensional Euclidean distances are converted into low dimensional conditional probabilities, but now using a student t-distribution. If similarity in high-dimensional space correctly models similarity in low-dimensional space, between any two points, the corresponding high- and low-dimensional conditional probabilities are equal. As a result of this, a new low-dimensional space that recapitulates the original relative arrangement or points can be derived from minimizing the disparity between both sets of conditional probabilities.(Maaten and Hinton, 2008)

Other well-known multi-dimensional scaling techniques such as principal components analysis (PCA) or SMACOF (more frequently known as classical MDS) are linear in nature and focus on maintaining large distances during the high–to-low dimensional space conversion. PCA, for example, is simply concerned with maximizing the variance of each

one-dimensional projection derived from the original data. Hence, such linear techniques lack the ability to keep small distances between the two spaces. This becomes a major shortcoming for high-dimensional data that lies on, or near low-dimensional, non-linear topological space. (Maaten and Hinton, 2008)

In this thesis t-SNE was employed, for the purpose of visualization, over a set of 1024-bit Morgan circular fingerprints (RDKit-equivalent of ECFP4), calculated for a radius of 2. To compute the t-SNE projection, an implementation in python, provided by the developer (https://lvdmaaten.github.io/tsne/#implementations), was used. The t-SNE projection was done using a perplexity of 30, an early exaggeration factor of 12 and a pre-processing step with Principal Components Analysis, where the top 50 dimensions are kept for the t-SNE calculation.

3.8. Project Workflow

The final endpoint pursued in this thesis is the prediction of the human volume of distribution. As established in the Introduction – Chapter 1, this is an endpoint of great relevance in a drug development context, but at the same time it entails a complex interplay of physiological processes. The study of the volume of distribution and its driving factors was carried through the main workflow shown in black arrows in Figure 3.3. In addition, a complementary workflow on applicability domain development was carried out, marked by the grey arrow.

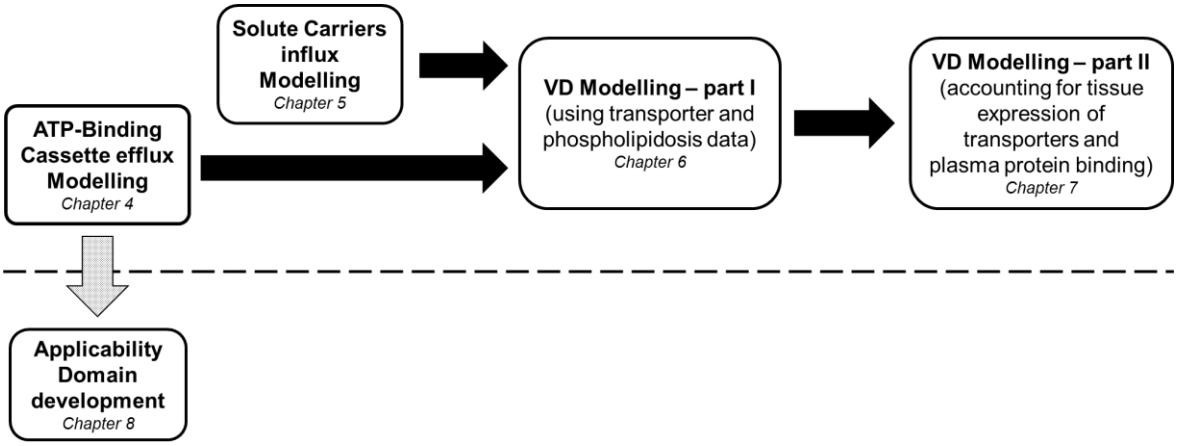


Figure 3.3. General outline of the thesis' workflow.

Workflow 1: Exploiting the Role of Physiological Data in the prediction of Vd

The main workflow is connected with arrows filled in black in Figure 3.1. This stream of work was organised to achieve a QSAR model for the prediction of Vd. For complex problems, such as this one, machine learning approaches can be effectively used to harness relevant information and uncover underlying processes which will help inform what drives the modelled endpoint in the first place. However, in order to do so, one needs informative input to feed into a machine learning algorithm. The classical approach adopted so far towards modelling volume of distribution via machine learning is mainly using easily attainable molecular descriptors, and despite some success having been achieved with some predictive models in the literature, this still remains a challenging endpoint to model for large and varied collections of compounds.

To directly tackle this issue in the attempt to improve predictive performance, as well as further the understanding of distribution and what drives it, this thesis will explore the hypothesis that adding input variables of physiological nature may improve the ability to model Vd. As presented earlier in this chapter, transporters are the main drivers of distribution, which makes them likely to be useful as input features. Additionally, drug-induced phospholipidosis and plasma protein binding are important determinants of Vd and have also been integrated as input in the Vd modelling process in this thesis. However, the relatively limited availability of experimental data renders this a prohibitive approach. Hence, in order to circumvent this issue some of the various physiological descriptors were modelled beforehand, and the learned responses output by the trained models were fed into the modelling of Vd.

Workflow 2: Exploring Applicability Domain Characterization of Pharmaceutical Data

In a complementary stream of work, this thesis will discuss the development of a novel applicability domain, which is one of the main current concerns in the field of chemoinformatics. Usually any given work focuses on either the modelling task or the applicability domain characterization task, however addressing both sides in the same body of work may provide importance context to each of them as, in fact, modelling and applicability domain characterization are conceptually co-dependent. Therefore, this part of the thesis will create a bridge between mispredictions (and what causes them) in their respective QSAR model and how they are perceived in terms of reliability in the established applicability domain.

The developed applicability domain algorithm will attempt to capture local properties of the data that determine reliability, namely density, bias and precision. Note that this will be the first attempt to harness information on bias associated with the training set to establish a

QSAR model's boundaries for reliability. This will be developed using the P-gp binding model as basis for the development of the algorithm, and will be benchmarked using two publicly available datasets used in the past for AD benchmarking (the Ames mutagenicity dataset and the CYP450 inhibition dataset).

3.9. Summary

This chapter described the details on datasets used, the calculation of descriptors used as independent variables, the employment of the different statistical methods for feature selection and machine learning. Additionally, the different evaluation metrics were provided, as well as details on applicability domain and activity cliff characterisation, and data visualization.

Finally an overview of the thesis workflow was described. Two main types of endpoints will be modelled in this thesis:

(1) Transporter data, where machine learning will be employed to differentiate between substrates and non-substrates for each of different ABC and SLC transporters. Multi-label methods will be used for the modelling of this data, which is motivated by the existence of overlap between different transporters;

(2) Volume of distribution data (reported at steady state, Vss), where information learned from the previous models will be fed into the machine learning algorithm as input features, thus allowing covering a wider range of data (as it cover missing observations in the data). A total of 12 physiological input features were used in the modelling of Vss drawn from transporter data, plasma protein binding data and phospholipidosis data. Additionally, transporter expression data will be used as a way to provide differential weight to the different input transporter features.

4. Multi-label Classification of ATP-Binding Cassette (ABC) Transporters

4.1. Introduction

As established in the Introduction chapter, ABCs are one of the main targets under focus in drug discovery and development, as major determinants of druggability given their potential to hamper absorption and distribution, as well as to potentiate excretion. Additional interest in this family of transporters comes from their role in multi-drug resistance in various cancer cells.

However, uncovering the underlying patterns that drive molecular recognition and subsequent efflux by ABC members still remains a challenge mainly due to the poly-specific nature of substrate recognition by these transporters. QSAR appears to be a particularly well suited method to predict ABC transport efflux, since it has been shown that substrate recognition relies on global physicochemical profiles rather than following the key-and-lock ligand binding model (Marquez and Bambeke, 2011). The potential of using QSAR to predict ABC transporter substrates during the R&D process has already been demonstrated by Desai et al. (Desai et al., 2013), who reported the successful replacement of an in vitro automated assay with a QSAR model to predict P-gp substrates in an early stage of the drug development pipeline of Eli Lilly.

Knowing that there is some degree of overlap between the binding patterns of different ABC members (Marquez and Bambeke, 2011, Wind and Holen, 2011), this can be exploited as a complementary source of information to aid the learning of the efflux process of different transporters. Multi-label classification is a suitable approach for this purpose as it accounts for overlapping information between different responses, which are addressed as a whole as opposed to the traditional single-label classification approach which looks at each transporter individually. The theoretical basis of multi-label classification has been detailed in the introductory Chapter 2 (section 2.6).

This chapter explores the creation of a multi-label QSAR of four major ABC transporters, namely BCRP1, MDR1/P-gp, MRP1, MRP2, with this being the first reported attempt to distinguish substrates from non-substrates of multiple ABC transporters using a multi-label classification approach. The goal of this study was to assess the potential value of taking

into account the data overlap amongst transporters in terms of the predictive accuracy of the classifier, as well as finding molecular characteristics that are unique to, or overlap between, the substrates of various transporters. The two previously mentioned multi-label classification schemes, Binary Relevance (BR) and Classifier Chain (CC), were employed; where the main difference between them is the presence of communication between transporter models, respectively.

A comprehensive validation routine including the characterization of the applicability domain (AD) and activity cliffs were carried out for the models. The predictive performance was analyzed against each model's applicability domain and activity cliff analysis, in the attempt of providing a more holistic, in-depth interpretation of the models' true worth. At the moment of publication, to the knowledge of the authors in the article, this is the first reported multi-label classification model for the prediction of ABC substrates and non-substrates, providing insight on transporter relationships with regard to binding patterns.

The contents of this chapter have been published in Molecular Informatics, under the following reference: Aniceto N, Freitas AA, Bender A, Ghafourian T: Simultaneous prediction of four ATP-binding cassette transporters substrates using multi-label QSAR. Molecular Informatics. 2016. 35. 514–28.

4.2. Methods

4.2.1. Dataset

Prior to any modelling or modelling-related task the ABC efflux dataset (described in Chapter 3) was submitted to a stratification procedure as described by Sechidis et al. (Sechidis et al., 2011). The authors show that this procedure leads to data subsets with more balanced class label distributions in a series of benchmark datasets. That is, this procedure maximizes transporters distribution across different data partitions. The stratification procedure was implemented in R using the provided pseudo-code by Sechidis et al. Consequently, the dataset was divided into training, (internal) validation and test set in a proportion of 3:1:1 (895 + 299 + 299 compounds), respectively, with similar distribution of substrates and non-substrates in all three subsets. Given the presence of a number of experimental conditions that can be optimized, in order to maximize the chance to select a set of conditions that retains ability to successfully predict unseen data, it is necessary that such optimization is guided by intermediate testing. Doing this with a subset of data outside

of the training set gives a more reliable assessment of performance when compared to using cross validation performance. This subset corresponded to the internal validation set. At the same time, to avoid overfitting and properly test the built model after all optimization has been done, a seconds split of the data should be reserved for final testing. This corresponded to the test set. In this work the test set was exclusively used to test the final models produced. As a future reference, the same principle was applied to all work carried in this thesis. The class imbalance across all four labels was deemed negligible, ranging between 1.0 and 1.7 (see Figure 3.1).

4.2.2. Molecular Descriptors

Molecular descriptors were calculated from SMILES codes retrieved from Metrabase (Mak et al., 2015). Structures were prepared for calculation following the specifications in the methodology section 3.2. In this work, molecular mechanics minimization was performed with the MMFF94x forcefield. The MOE version used was v 2013. A total of 338 molecular descriptors were obtained and submitted to feature selection.

4.2.3. Feature Selection

Even though the C4.5 algorithm incorporates its own embedded feature selection (see section 2.5.1), it has been reported to overfit, producing very large trees. In order to minimize the risk of overfitting it is recommended that feature selection is employed prior to training (Kohavi and John, 1997).

A total of five feature sets derived from the five different feature selection techniques (Genetic Algorithm Search, GA; Greedy Stepwise Search, GS; ReliefF, RfF; decision trees+Genetic Algorithm Search, C4.5-GA; and Random Forest-Greedy Stepwise search,,RF-GS) described in Chapter 3 were produced for each of the four ABC transporters' subsets. These resulted from ranking all available features and taking the top 20 top ranked features (for tied ranking features within the top 20 threshold value were also included). All calculations were done using WEKA 3.6. Each of the feature sets was subsequently used to train a C4.5 model, for each of the labels (transporters). In order to select the best feature selection method for each transporter label, C4.5 models were trained with the different feature sets. These models were then tested on an independent internal validation data subset. This corresponds to a total of 20 experiments testing five different feature sets for each of the four ABC transporters. The best feature set for each transporter was selected according to the highest Matthews correlation coefficient (MCC) and geometric mean between sensitivity and specificity (G-mean) in the internal validation set (Zhou et al., 2015), both defined by Equations 3.3 and 3.4, respectively. A summary of

performance obtained for all feature selection + transporter is provided in Appendix I (Section 11.1).

4.2.4. Multi-label QSAR models

The best C4.5 models (using the best feature selection conditions) produced from the feature selection optimization step were selected for each of the training sets (BCRP1=288, MDR1=580, MRP1=111, MRP2=145).

The multi-label Binary Relevance (BR) model was obtained by gathering the predictions from these four best single-label models into one global prediction output. In this case, whenever a new query compound needs to be predicted it would be passed through all four ABC models and a set of label predictions would be produced. For the multi-label classifier chain (CC) model, the schematic representation of CC is depicted in Figure 4.1. The transporters were ordered according to descending order of dataset size, based on the theoretical expectation that larger datasets will have a better chance of providing useful information to smaller datasets than the other way around. Accordingly, the order of the labels in the classifier chain was P-gp/MDR1 > BCRP1 > MRP2 > MRP1. To build the multi-label CC model each label (transporter) in the 4-label chain uses the best descriptor set previously optimized for the BR model. In addition, as it can be seen in Figure 4.1, each label in the CC model uses prediction sets from previously available labels. In summary, in the CC model every label (transporter) in the chain is trained using the prediction sets from all previous labels, along with a set of molecular descriptors (previously selected). To illustrate this, label #3 for example, will be trained with a set of molecular descriptors as well as class predictions for label #1 and #2.

Overall each transporter was submitted to an independent and parallel process of feature selection, model optimization and training, and finally testing. All these steps were performed in parallel on the same datasets for CC and BR in order to: 1) allow comparability between both types of model at every level, and 2) assess the value of addressing the overlap in the data, by fixing all other conditions in both modelling workflows. Throughout the paper the following notation <single-label model> - <multi-label model> will be used whenever a specific single-label model within the CC or the BR models is mentioned.

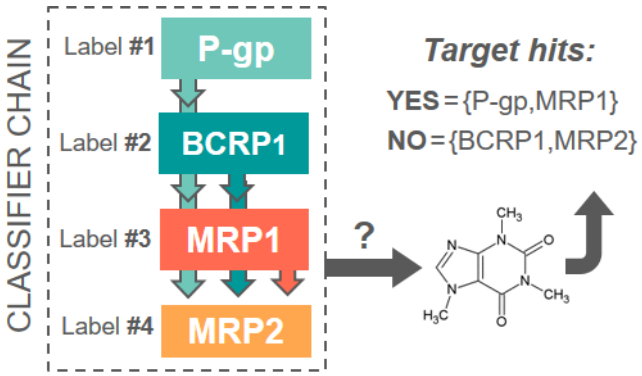


Figure 4.1. Schematic representation of multi-label classifier chain training.

4.2.5. Model Validation

Both BR and CC multi-label models were assessed and compared for their predictive performance through the various measures provided in section 3.5. This evaluation was done at the multi-label level, and at the single-label level as well (by looking at each label's performance within a given multi-label model).

Additionally, the models were assessed regarding their applicability domain (AD). The AD of all the single label models used in the generation of multi-label BR and CC models was characterized, by using STD scores (calculated as defined in the Methods section) as a measure of predictive reliability. As a complement to the AD analysis, potentially relevant activity cliffs were identified, by locating compounds of high similarity (structurally) but with opposite responses. Similarity was measured as the Tanimoto coefficient of Morgan fingerprints, and the threshold for significant Tanimoto coefficient was set to a minimum of 0.55.

To assess the chemical space of the ABC efflux dataset with relation to the real-world drug chemical space, the ABC transporter data was overlaid against the DrugBank chemical space. This was done using t-SNE multidimentional scaling performed on the full DrugBank data and the ABC efflux data.

4.3. Results and Discussion

4.3.1. Multi-label QSAR models

In this work, the main goal was to model four ABC transporters in such a way that allows accounting for possible underlying correlations between labels (i.e. transporters). Multi-label classification is the appropriate approach to achieve this. The multi-label models were built

using a decision tree learner (C4.5), as this machine learning algorithm has a visual and transparent nature that allows interpretation of the effects of the features on the predicted labels. Furthermore, decision trees can cope with different scales in the descriptors and they can also handle both continuous and categorical data efficiently and robustly (Dehmer and Varmuza, 2012).

By comparing a multi-label method that accounts for label interaction (i.e., CC) with an alternative method that assumes labels to be independent (i.e. BR) one is able determine whether label interaction (i.e. correlation between the binding profiles), in fact, exists among the different ABCs. Both multi-label classifiers were trained using the best features selected by various feature selection methods for each transporter, and they differ only in the use of previous label predictions as additional features (in the case of CC). The rational for the use of multi-label methods was the overlap observed in the dataset, as can be seen from the results of the Chi-squared test measuring the correlations between labels (Table 4.1). These multi-label methods were compared in terms of their predictive accuracy in the classification of various ABC transporters' substrates and non-substrates.

Table 4.1. Values of the Chi-squared test measuring correlation between labels. The smaller the Chi-squared value, the stronger the chance of true correlation.

	MDR1	**MRP1**	**MRP2**
BCRP1	**0.001**	**0.001**	**<0.001**
MDR1		**<0.001**	0.679
MRP1			**<0.001**

Within each multi-label model it is necessary to make sure that each one of its single-label models provides a reasonable input to the global multi-label model. Firstly, the best single-label C4.5 model for each transporter was selected out of a pool of five models obtained from various pre-processing feature selection methods. The results showed that the GS method led to the best model for BCPR1, while C4.5-GA led to the best models for MDR1 and MRP1; and ReliefF led to the best model for MRP2 (Appendix I, Table A1.1). Table 4.2 shows the performance of the best single-label models. Secondly, to validate the inclusion of each label, the impact of removing or adding a label on the overall performance of BR and CC models, respectively, was assessed using the Hamming Loss, with respect to the internal validation set (Figure 4.2). Note that the Hamming Loss measure ignores interaction between labels, since its value depends on whether or not each label was correctly predicted by itself, regardless of the predictions of the other labels. Both BR and CC models show a constant impact in Hamming Loss by the presence of all labels, which is depicted by a constant Hamming Loss value as the chain grows, in the CC, and when different labels are removed in turn, in BR (Figure 4.2). This observation justifies the presence of each label

in the multi-label models. The same is observed in the test set where no particular label stood out in terms of impact on Hamming Loss performance (Figure 4.3) which means no label is causing degradation of the predictive performance.

Table 4.2. Test set performance of the single-label models for individual transporters using the best set of features with (CC) or without (BR) the use of the predicted ABC binding class of the preceding transporters in the classifier chain. Values expressed in percentage.

	MDR1 (n=195)	BCRP1 (n=87)		MRP2 (n=41)		MRP1 (n=36)	
	C4.5-GA	GS	GS pMDR1	RfF	RfF pMDR1 pBCRP1	C4.5-GA	C4.5-GA pMDR1 pBCRP1 pMRP2
G-mean	66.8	76.3	**76.7**	74.4	74.4	58.9	**59.0**
Sen	79.1	84.5	77.6	69.2	69.2	84.2	74.0
Spe	56.5	69.0	75.9	80.0	80.0	41.2	47.1
MCC	36.6	53.4	51.4	47.4	47.4	28.3	21.6

At the multi-label level, Table 4.3 indicates a good performance with an overall F1 of approximately 70% for both BR and CC models. Even though this represents a 30% error rate, it is considered a good prediction performance in light of how challenging it is to model the current data. This is due to the large level of noise and ambiguity in the data, which is discussed later in the discussed section, as well as in light of the imbalance of information provided for substrates versus non-substrates. The results also show that both models performed very similarly. However, attention must be drawn to the fact that the modelled data is imbalanced both at the label level (i.e. some transporters have a higher proportion of substrates than others) and, to a lesser extent, at the class level (i.e. within each transporter there is more substrates than non-substrates). This means that commonly employed measures, such as F1, precision and recall, will be leveraged by the majority label and the majority class, and therefore they are not ideal to assess these imbalanced problems. Alternatively, balanced accuracy (bACC) is designed to overcome this issue. Table 4.3 shows that bACC has a higher score for CC. Additionally the CC model shows less discrepancy between the ability to predict substrates and non-substrates, shown by the absolute difference between precision and recall (ΔPR). This means the CC model achieves the best balance in terms of classifying both substrates and non-substrates.

Moreover, a comparison of single-label (individual transporter) models used to develop BR and CC (Table 4.2) shows that the two single-label models that include a predicted label as a feature (BCRP1, and MRP1) have improved Sen-to-Spe balance (values highlighted in gray), which supports the existence of label correlations and the advantage of taking them

into account when modelling ABC transport data by using CC instead of BR. The model for MRP2 stayed at the same level of accuracy, which means that information from other transporters used in the CC model (left-hand side) recapitulates information from chemical nature used in the BR model (right-hand side).

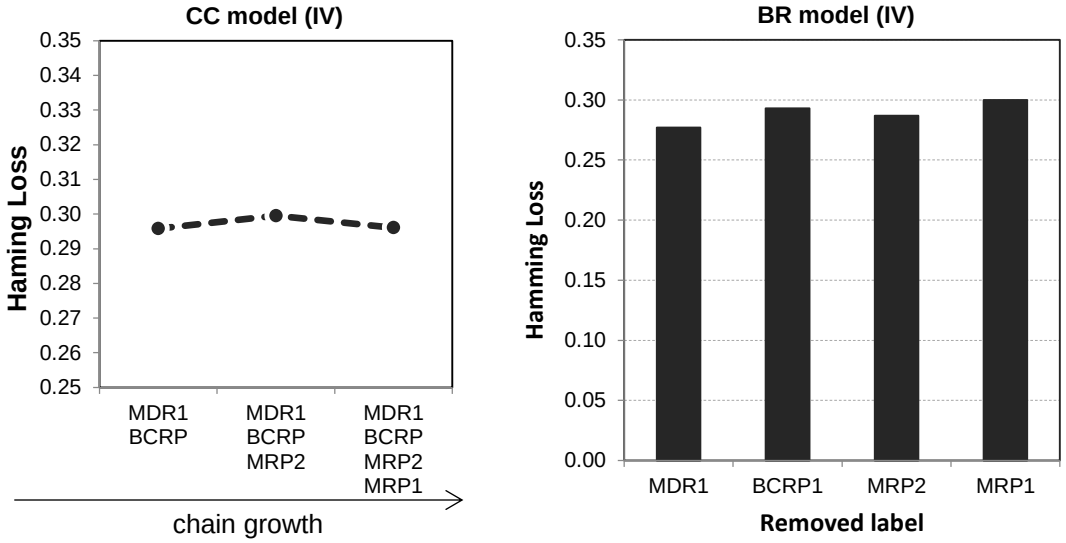


Figure 4.2. Impact of each label on the overall performance of the CC and BR models, tested on the internal validation set. The graph for CC depicts the evolution of the model's performance as labels are being added to the chain, whereas the graph for BR depicts the model's performance when each of the labels is removed, in turn.

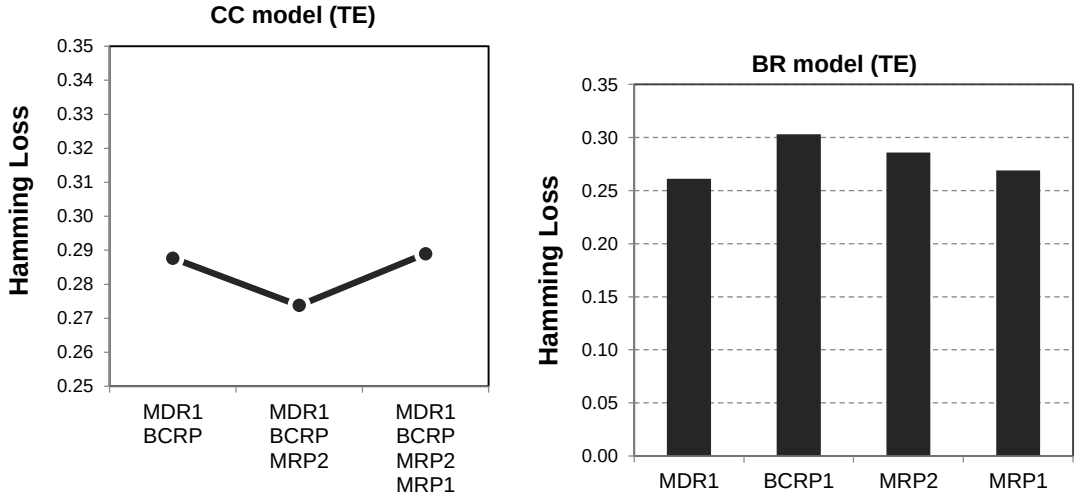


Figure 4.3. Impact of each label on the overall predictive test performance of the CC and BR models. The graph for CC depicts the evolution of the model's performance as labels are being added to the chain, whereas the graph for BR depicts the model's performance when each of the labels is removed, in turn.

Table 4.3. Summary of performance measures of the final BR and CC models in the test set. Underlined font marks the values that are better than their direct counterpart models.

Performance measures	Equation used	BR	CC
F1	3.6	69.6 %	69.2 %
bACC	3.9	68.7 %	69.0 %
Precision	3.7	70.4 %	70.0 %
Recall	3.8	70.0 %	69.6 %
ΔPR	3.10	20.6 %	17.4 %

Overall, despite the substrate overlap between various members of the ABC transporters mentioned earlier and by others in the literature (Vastag et al., 2011, Bentz et al., 2013), BR and CC yielded very similar predictive performance statistics. On the other hand, it is apparent that the predicted MDR1 class (pMDR1) is favored over molecular descriptors in the BCRP1 model, and the predicted MRP2 class (pMRP2) is preferred in the MRP1 model, as evidenced by the preferential selection of these features as one of the top five model features (compare BR and CC features in Table 4.4 and Table 4.5). There are several possible explanations for the lack of a significant improvement of CC comparatively to BR (Table 4.3). The first possible explanation may be that labels have close to no interaction, which means that the classifier chain has nothing to capitalize from. However, Table 4.1 shows that all pairs of labels, except one, have a significant correlation, so the issue with regard to this hypothesis may be instead the relatively low label density (the compound vs label matrix is only 23% populated in the training set), which reveals scarcity of multi-label cases (i.e., compounds with measured binding in several transporter systems). The second possible explanation may the fact that the BR model depends on the individual quality of each single-label model; while the quality of the CC model depends also on the quality of the prediction of the previous labels in the chain. In fact, in a CC model every flaw in any given label (transporter) will be carried on to the following labels in the chain, as opposed to BR, in which the shortcomings of a model have no effect on the remaining labels.

Even though the final overall statistics show no marked numerical improvement from accounting for label interaction, focusing only on this can give an overly simplistic view. When results are analyzed as a whole, there are several pieces of evidence of the value of using label interaction in the modelling of the ABC QSAR. In two of the three single-label models, built by the CC method, where previous labels were available, previous label information was spontaneously selected by the decision tree building algorithm. Furthermore, this singular change in the entire modelling process coincided with more parsimonious models, which showed more balanced Sen to Spe ratio. This is a very

valuable improvement given that this modelling task would naturally tend towards higher Sen, brought on by the relatively larger amount of substrates than non-substrates. Class imbalance within each label is known to have yielded poor models in the past (Sedykh et al., 2013, Newby et al., 2013), and being able to mitigate this issue without using any type of aiding technique (i.e., over-/under-sampling or misclassification cost) is notable.

4.3.2. Molecular Descriptors in Single-label Elements of BR and CC

As it was explained in previous sections, the molecular descriptors used in C4.5 models have been selected by the best pre-processing feature selection methods for each transporter dataset followed by the embedded C4.5 feature selection. Among the five feature selection methods, C4.5-GA features yielded the best results for the majority of single-label models. The purpose of using a wrapper rather than a filter method is to select a feature set that ideally best copes with the classification algorithm's biases. However, given the complex nature of these transporters it is expected that different feature-selection methods are best suited for the predictions of different labels, and indeed this has been observed in the results.

Roughly the same number of molecular descriptors was provided to the C4.5 algorithm for the modelling of each transporter, however the number of descriptors used to build each tree decreased along the order of the labels in the chain, i.e. MDR1, BCRP1, MRP2, and MRP1. Moreover, recall that the same set of molecular descriptors was provided to C4.5 for the single-label constituents of the BR and CC models, but single-label elements of CC employ additional predictors, i.e. the predicted label (substrate vs. non-substrate) of the previous transporter(s) in the chain.

Given the large number of molecular descriptors incorporated in some C4.5 models, these descriptors can be ranked according to their statistical importance and the most important molecular descriptors may be identified. Tables 4.4 and 4.5 show the importance of molecular descriptors in C4.5 models for different transporters in BR and CC models, respectively. These molecular descriptors have been described in Appendix I, A1.2. In order to calculate the feature importance, the molecular descriptors used in the models were ranked according to the number of compounds that were directly affected by each descriptor at any point of the tree. In this way, descriptors selected earlier on for nodes closer to the root of the trees are more important than those selected later on (closer to leaf nodes) to classify a smaller number of compounds. Table 4.5 shows that the molecular descriptors selected by the C4.5 algorithm for BCRP1 and MRP1 include a transporter substrate class predicted by the previous transporters in the chain, and both predicted labels used in both models affected more than 50% of the training data (see Table 4.5).

Due to the design of the CC model that placed the MDR1 transporter as the first label in the chain, the single-label MDR1 model used in both multi-label BR and CC models is the same, i.e. no predicted ABC label was used as a feature in the modelling of this transporter. As a result, the MDR1 descriptors reported in Tables 4.4 and 4.5 are the same. For BCRP1, a comparison of Tables 4.4 and 4.5 shows that some of the molecular descriptors in the BR model have been replaced by the predicted MDR1 class as an important feature in the CC model of BCRP1. On the other hand, the single label MRP2 model developed by C4.5 did not pick predicted MDR1 or predicted BCRP1 labels, and only molecular descriptors were selected as the model features. As a result, the top descriptors used in the single label MRP2 models within both BR and CC models are the same (see Table 4.4 and 4.5). For MRP1 models, a comparison of Tables 4.4 and 4.5 shows that the models developed for CC and BR are different, as the predicted MRP2 labels have been used in the multi-label MRP1 model built by the CC model. The MRP1 model for CC used the predicted MRP2 label as the second most important feature replacing the polar volume.

Common features between transporters could be an indication of the degree of shared substrates. MDR1 and MRP1 both share the same best feature selection method (C4.5-GA) and there is some degree of feature overlap (around 5 features) between them. MDR1 shows the strongest correlation with MRP1 (Chi-squared test, $p < 0.001$, Table 4.1), and in fact there is a considerable amount of common substrates and non-substrates between them (n=34 and n=12, respectively out of 61 common compounds). The overlap of substrates between various ABC transporters is a well-established phenomenon (Matsson et al., 2009). For instance, it was reported that drug resistance to daunorubicin derives from a synergy between MRP1 and MDR1 activities (Legrand et al., 1999).

The nature of the molecular descriptors incorporated into the single label C4.5 models can provide clues for the molecular characteristics of a compound associated to molecular recognition by a transporter as its substrate (See the Appendix I, Table A1.2). Extending this reasoning to the multi-label perspective, looking at the composition of the decision tree models, molecular descriptors show some overlap between different transporter models, which supports the multi-label approach from a mechanistic standpoint. In particular, features of MDR1 and BCRP1 substrates have some similarity as both transporter's substrates are bulky and flexible, and contain hydrophobic moieties. MDR1 substrates are highly branched, good electron acceptors (such as in hydrogen bonds) and contain quaternary ammoniums, while BCRP1 substrates contain large positively charged surface, have aromatic rings and may be a non-drug-like molecule. The correlation of these two transporters is evidenced by the fact that the predicted MDR1 label is a very useful feature for the classification of BCRP1 transport.

Table 4.4. Descriptor importance calculated from the relative amount (%N) of compounds classified using every given feature within the BR model. Predicted labels are suffixed with the feature set that originated them. See Appendix I, A1.2 for descriptor definitions.

MDR1 (C4.5-GA)	%N	BCRP1 (GS)	%N	MRP2 (RfF)	%N	MRP1 (C4.5-GA)	%N
VDistMa	100	Num_Rings_4	100	ast_violation_ext	100	Q_VSA_POL	100
FCharge	85	Q_VSA_FPPOS	94	PEOE_VSA_FPNEG	65	vsurf_Wp1	70
a_nH	80	SlogP_VSA7	82	vsurf_CW2	61	Q_VSA_FPPOS	53
b_max1len	64	b_ar	68	reactive	54	FCASA+	38
PM3_LUMO	63	opr_nring	53	Fi(B)	34	chi1v_C	34
PEOE_VSA+6	52	a_nF	30	b_rotR	24	b_rotR	30
SMR_VSA2	45	glob	24	opr_leadlike	16	b_max1len	15
a_acc	27	a_ICM	23	Q_VSA_FHYD	12	Kier3	14
b_ar	25	PEOE_VSA-3	22	vsurf_HB2	11		
dens	22	LogD(6.5)	19	Fi(A)	4		
PEOE_VSA-6	20	MNDO_LUMO	18				
Num_Rings_5	16	SMR_VSA4	9				
FCASA-	13	LogD(5.5)	5				
vsurf_Wp5	11	PEOE_VSA-4	3				
vsurf_Wp6	10	PEOE_VSA-1	2				
SlogP	8	vsurf_R	2				
Rule_Of_5	8	LogD(7.4)	2				
PM3_E	8						
MW	3						
vsurf_CW8	2						
PEOE_VSA_NEG	2						
Polarizability	2						

On the other hand, molecular features of MRP2 and MRP1 substrates are also similar in terms of polarity and hydrophilicity of the molecular surface. MRP2 substrates may contain reactive groups defined as nitrogen, oxygen and sulfur atoms with polar negative surface area, while MRP2 substrates are flexible in addition to large polar and hydrophilic surface area. Furthermore, the predicted MRP2 binding class can be used as a significant feature for the prediction of MRP1 transport. MDR1 and BCRP1 were more associated with explicit aromaticity-related features, whereas MRP1 and MRP2 where predominately more

associated with hydrophilicity-related properties, which could be tied with the fact that MDR1 and MRP2 were used as predictors in both BCRP1 and MRP1 models respectively.

Table 4.5. Descriptor importance calculated from the amount of compounds classified using every given feature within the CC model. Predicted labels are suffixed with the feature set that originated them. See Appendix I, A1.2 for descriptor definitions.

MDR1 (C4.5-GA)	%N	BCRP1 (GS)	%N	MRP2 (ReliefF)	%N	MRP1 (C4.5-GA)	%N
VDistMa	100	Num_Rings_4	100	ast_violation_ext	100	Q_VSA_POL	100
FCharge	85	Q_VSA_FPPOS	94	PEOE_VSA_FPNEG	65	pMRP2_ReliefF	70
a_nH	80	SlogP_VSA7	82	vsurf_CW2	61	vsurf_D7	46
b_max1len	64	b_ar	62	reactive	54	b_rotR	30
PM3_LUMO	63	pMDR1_C4.5-GA	55	Fi(B)	34	Q_VSA_FPPOS	24
PEOE_VSA+6	52	opr_nring	48	b_rotR	24	rings	17
SMR_VSA2	45	glob	46	Q_VSA_FHYD	12	b_max1len	14
a_acc	27	a_nF	30	vsurf_HB2	11		
b_ar	25	PEOE_VSA-3	22				
dens	22	MNDO_LUMO	21				
PEOE_VSA-6	20	vsurf_CW2	19				
Num_Rings_5	16	LogD(6.5)	19				
FCASA-	13	a_ICM	9				
vsurf_Wp5	11	LogD(5.5)	7				
vsurf_Wp6	10	SMR_VSA4	7				
SlogP	8	a_aro	4				
Rule_Of_5	8	PEOE_VSA-4	3				
PM3_E	8	vsurf_R	2				
MW	3	LogD(7.4)	2				
vsurf_CW8	2						
PEOE_VSA_NEG	2						
Polarizability	2						

4.3.3. Applicability Domain and Activity Cliffs

By applying the STD method as per Sushko et al. (Sushko et al., 2010a), it is possible to observe an overall declining trend of accuracy as a function of STD, across the majority of the single-label models (Figure 4.4). Exceptions to this trend will be further explored.

There are two main important aspects to consider for the quality of an AD profile, similarity of overall profiles/trends (i.e. similar slope direction) and a decreasing accuracy as the chemical space moves away from the model's core. Exploring Figure 4.4 points to only two cases where the requirements above have not been met; these are MDR1-BR and MRP2-BR in the validation set. This is not seen for the corresponding CC model, MRP2-CC, (note that the MDR1 single-label model is the same in both BR and CC models). There is also a mild case of disparity between validation and tests sets for BCRP1 (although only at the first iteration of STD increments). While this disparity happens for BR, in the CC model all trends start at a higher point and tend to decrease with STD (although this is not done in a perfectly smooth way, as expected from any kind of AD analysis).

Interestingly, even though MRP2 models show the exact same performance statistics at the single-label level (Table 4.2), there is a marked difference between the AD profiles of its BR and CC single-label models developed using a 10-fold bagging ensemble, depicted in Figure 4.4. This is evidence that the presence of previous labels allowed establishing a more reliable AD of the model. Even though both MRP2-CC and MRP2-BR yielded equal predictive performance, MRP2-CC allows a better definition of its applicability as both external datasets show the same trend of accuracy vs STD (Figure 4.4). As the AD method is insensitive to bias and relies solely on precision, low STD scores may happen due to a systematic misprediction in all models in the ensemble rather than a reliable (correct) prediction. This systematic misprediction in the low STD area was the case in MRP2-BR. On the other hand, the presence of two extra features in MRP2-CC (the two previous labels in the chain), which were picked for 3 of the 10 bagged models, helped overcome the systematic bias in modelling MRP2 data. Therefore, MRP2-CC allows establishing a threshold of prediction reliability that imitates the reliability trend in external data. As a result, these observations consist of a proof of concept of the value of using CC for the purpose of modelling ABC substrate data.

Lastly, it should be noted that, for some labels, the increase in accuracy is not significant for smaller STD values. This is due to the quality of the trained model that may not allow a high level of precision (agreement between the ensemble models). Still, even if there is a small gain in accuracy at a given threshold, this still entails a decreased risk of producing a wrong prediction, and thus the respective AD profile is useful in guiding the prediction acceptance.

Even though this analysis gives insight into a model's overall predictive performance across the data, it is convenient to further pinpoint activity cliff regions. To this end, activity cliff analysis was used in this study to identify areas of high complexity in the structure-activity data. Table 4.6 shows that a considerable portion of activity cliffs coincides with mispredictions. These can be areas of higher complexity in terms of the structure-property

relationship that require more compounds and/or better use of molecular descriptors that would capture some subtle chemical variation (Maggiora, 2006). These can also result from unreliable experimental data (i.e., if a substrate is incorrectly presented to the learning algorithm as a non-substrate, even if it is correctly predicted as substrate it will be perceived as a misprediction) (Sedykh et al., 2013).

Recall that three single-label models in the multi-label classifier chain could use previous labels as descriptors (considering that MDR1, as the first label of the chain, cannot use previous label descriptors). The fact that in two out of those three models a considerable portion of the activity cliffs was associated with mispredictions shows the correlation between both. It should be pointed out that in both BCRP1 models (produced by the BR and CC methods) there were two compounds that were mispredicted in the former model while being correctly predicted in the latter.

Table 4.6. Comparison between activity cliffs (ACs) and mispredictions within them – values in brackets are the percentage of activity cliff compounds that are mispredicted by the models.

Transporter model	Number of ACs mispredicted	Number of ACs
MDR1 (BR/CC)	9 (50%)	18
BCRP1 (BR &CC)	4 (40%)[a]	10
MRP1 (BR & CC)	2 (100%)	2
MRP2 (BR & CC)	0	2

As an example, Figure 4.5 depicts the distribution of mispredictions (false negatives and false positives) for the BCRP1 BR model overlaid with the substrates and non-substrates. It can be seen that activity cliffs are mainly located in areas of sparse data especially at the extremities of the plot.

Mispredictions were further analyzed for their distribution along the test set chemical span of each of the molecular descriptors used in the various decision trees (all distribution graphs are shown in Appendix I, Figures A1.1-7). For all models in BR and CC, mispredictions overlap with correct predictions in the test set. Furthermore, it is common to find both mispredicted compounds close to the center-values, and correctly predicted compounds near data limits (and even outside the training range).

The validation and test sets were also analyzed for their distribution with respect to the training chemical span. This revealed no apparent trend in terms of misprediction

concentration in chemical space, with the mispredicted compounds often showing scattering centered at the median of each descriptor. As a matter of fact, mispredicted compounds seem to follow the distribution of the training set, being more densely located near the median and scattering away from it in a somewhat parallel manner. Additionally, both in MDR1 and BCRP1 datasets, despite some compounds being clear outliers with respect to certain individual descriptors, as seen in Appendix I, Figures A1.1-7, falling outside the maximum range of the training set ([0;1], standardized data) they were successfully predicted by their respective models. However, these observations were exceptions and, overall, the validation sets were found within the maximum range of each descriptor in the training set.

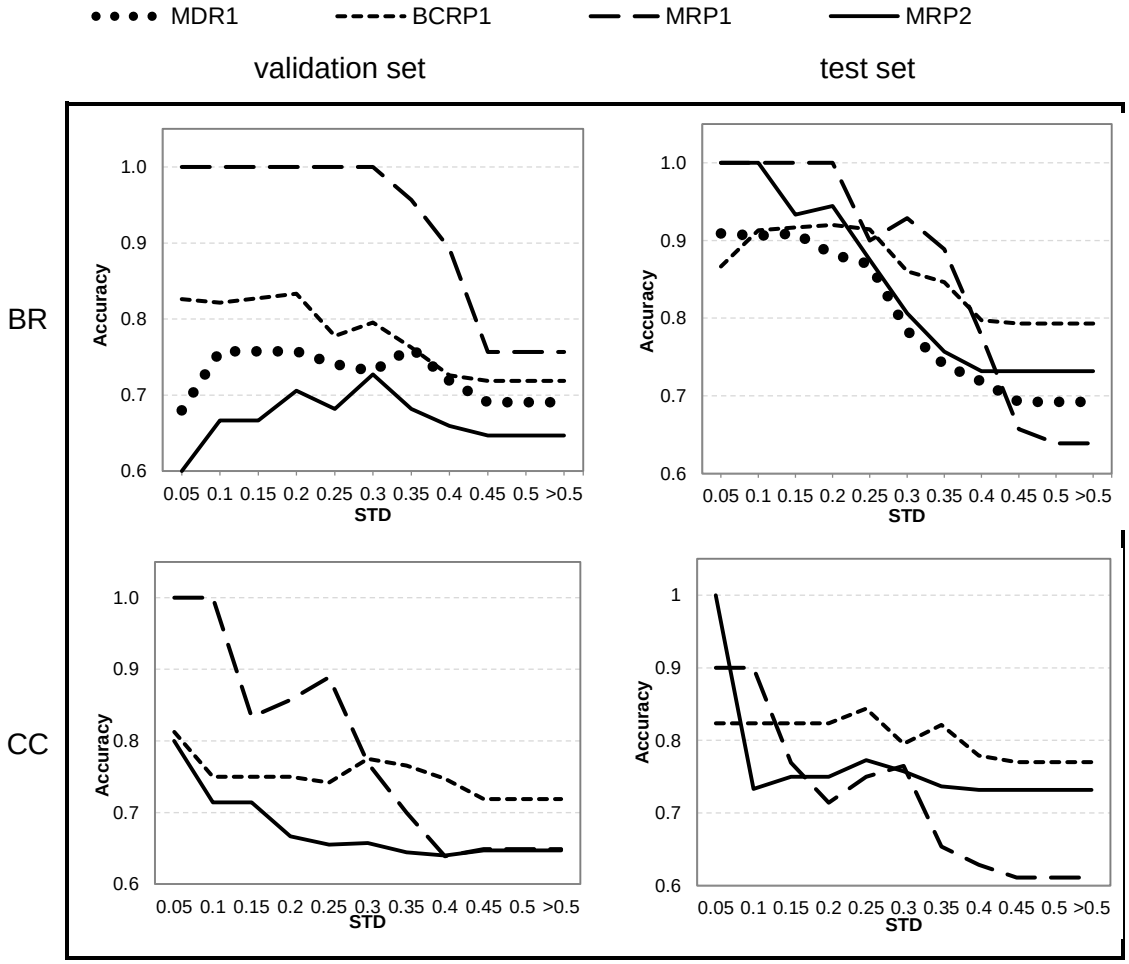


Figure 4.4. Applicability domain evaluated with respect to the validation and test sets. Recall that accuracy has been defined as the fraction of correct predictions out of the total number of predictions that fall within any given threshold (set in the axis labeled “STD”).

Apart from the applicability domain and activity cliff analysis, it is useful to analyze the range of chemical diversity covered by the models built, in order to support the validity of their

future predictions. This was achieved by overlaying the data from the four transporters with the DrugBank dataset using a t-SNE multidimentional scalling projection of the Euclidean distances (Figure 4.6). Considering that DrugBank holds the full span of chemical variety in real-world drug space, this analysis provides a gauge of the diversity in the data used in this work. Despite the scarcity of data in some transporter datasets, they were all evenly spread across the chemical space of the entire DrugBank dataset (more than 6000 instances). This means that the models incorporate a wide chemical variety in the training, which strengthens their potential usefulness as a predictive tool.

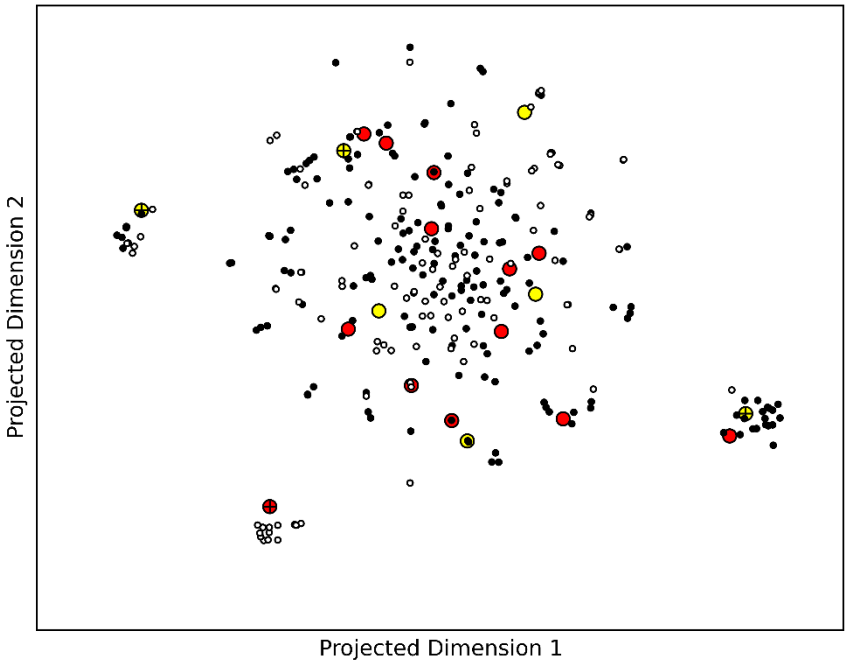


Figure 4.5. Mispredictions and activity cliffs of the BCRP1-BR model; Training data were projected into a 2D map using t-SNE, and the location reflects the Euclidean distance between ECFP4 fingerprints. The Tanimoto coefficient was not used as a visualization measure as it produces plots with very distant points. However, using the Euclidean distance conserves visually the relative neighborhood of each point. activity cliffs are marked with a cross; FP: yellow; FN: red; training substrates: black; training non-substrates: white.

The performance of the models developed in this work has to be evaluated in light of the high level of noise in any kind of large transporter dataset. Several factors are known to contribute to the considerable inter-laboratory and even inter-experimental variability in permeability/efflux assays. Some frequently reported examples are sensitivity to varied culture protocols and conditions, genetic variations of MDR1 (and other transporters) leading to variable pump functionality, and variable expression levels of various ABC transporters and even different additional transporters (i.e. Solute Carriers) (Ganta et al., 2008, Vastag et al., 2011). There are also parallel metabolizing enzymes and alternative

active transport systems. The variability is therefore a significant factor within a single dataset built from different sources using different cell models (Bentz et al., 2013). As a result, the BR and CC models should be evaluated in light of realistic maximum obtainable performance. In an ideal scenario a perfect model would correctly classify 100% of unambiguous cases (correctly belonging to their assigned classes), and would correctly classify 50% of ambiguous cases (given that probabilistically only 50% are actually correctly classified to begin with).

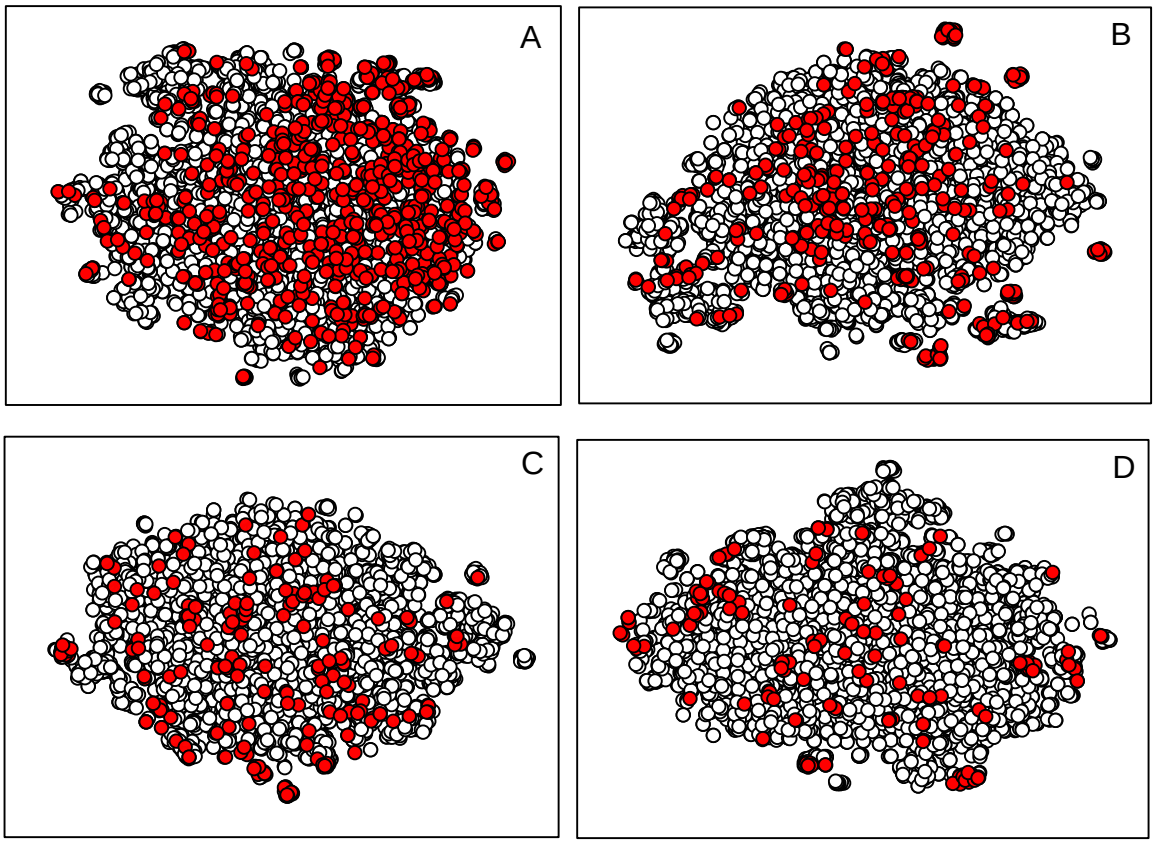


Figure 4.6. Chemical space coverage of MDR1/P-gp (A), BCRP1 (B), MRP2 (C) and MRP1 (D) with respect to the DrugBank complete dataset. The ABC datasets are represented in red in their respective scatterplots, and DrugBank data is depicted in white. The plots result from a t-SNE multidimentional scalling projection of ECFP4 fingerprints, and the two axes are projected dimension 1 and 2 obtained from the t-SNE embedding.

Applying this reasoning to this work's dataset translates into a maximum accuracy of 98% since the dataset has 61 ambiguous responses (i.e. reported as substrate and non-substrate from different sources) across 1493 compounds, hence 2% will theoretically be mispredicted. However, this is a conservative estimate, due to the inter-laboratory variations affecting the accuracy of a given label in the literature, where the majority of compounds in the dataset have only one experimental measurement**.** It must be noted that in the construction of Metrabase, the allocation of substrate and non-substrate labels was carried

based solely on the recommendation of the original literature reference (Mak et al., 2015). However, different literature sources have differing criteria and threshold values (in addition to varying experimental techniques) for classifying a compound as substrate (Broccatelli, 2012). A threshold of 2 for the efflux ratio is normally used by researchers, while the borderline interval is [1.8-2.5] (Broccatelli, 2012). In fact a maximum accuracy of 86% has been reported for MDR1 efflux assays (Broccatelli, 2012). In an overall appreciation of the feasibility of using the models presented here, as a substitute of the gold standard cell assays, these models are able to produce valid predictions in 70% of the cases, while the Borst cell assay (n=91, see Broccatelli et al (Broccatelli, 2012)) produced usable prediction in 76% of the cases considering that contradictory replicates (n=16) and borderline values (n=6) cannot be used to trustfully classify a given compound.

In this study, even for models that were trained on datasets with balanced classes, the specificity is always considerably lower than the sensitivity, which means that the models are generally more capable of identifying substrates than non-substrates. However, this is not unprecedented as several other works on MDR1 substrate prediction listed in the literature (Broccatelli, 2012) have reported the same issue. Comparing the results of two previous works where efflux ratios of 2 (Broccatelli, 2012) vs 2.5 (Gupta et al., 2010) have been used as threshold values, models with higher threshold values generally lead to lower specificity as expected. It can be hypothesized that the main underlying cause for a tendency for poor Spe is the fact that some substrates also have high passive permeability. This leads to cases of substrates that cannot be identified by permeability measurement methods (falsely identified as a non-substrates), which will translate into spurious data in the non-substrate class (Broccatelli, 2012).

To contextualize the potential utility of the CC model proposed here, as of 2012, Tsaioun and Kates (Tsaioun and Kates, 2012) reported a 15% increase in phase 2 failures, 50% of which are due to lack of efficacy. However, many of these failures are CNS-targeted clinical trials where lack of efficacy is caused by an underlying failure to permeate the BBB. It is safe to say that, considering the polyspecificity of MDR1 in addition to the presence of a large variety of other ABC exporters on the BBB, a large portion of this attrition rate could probably be associated to some extent with the efflux of the drugs in question. In fact, in retrospect it is possible to identify cases where, if this work's models had been used, it would have been possible to avoid very expensive clinical trials through the prediction of the substrate ability of different ABC substrates. Two examples from the test set are sunitinib and dasatinib, both predicted as MDR1 and BCRP1 substrates based on CC and BR models. Sunitinib failed a phase II clinical trial (NCT00923117) for the treatment of glioblastoma due to lack of efficacy. The probable cause for such late failure was that this drug has poor ability to permeate the BBB, which is most likely due to MDR1 and BCRP1

efflux (Oberoi et al., 2013). In retrospect, if the models herein developed had been applied to sunitinib, it would have been possible to avoid a failed clinical trial, since both BR and CC were able to predict this compound as a substrate of both transporters. Even if the trial was carried out, the use of a predictive model like the one reported here would at the least maximize the chances of success with the concomitant administration of an inhibitor. A similar scenario was observed for dasatinib, which showed no effectivity in a clinical study with 14 patients (Lu-Emerson et al., 2011).

4.4. Conclusions

This chapter reports two multi-label models for the prediction of various ABC transporter substrates and non-substrates, namely BCRP1, MDR1/P-gp, MRP1 and MRP2. The multi-label classifier chain (CC) method, which accounts for label (transporter) interaction, was compared with the binary relevance (BR) method, which does not consider interaction. Both models showed good predictive power, as expressed by F1 values (weighted average of precision and recall) and a balanced accuracy of approximately 70%. Even though the CC model showed no marked improvement in terms of the general performance measures, a closer analysis revealed several evidences of the benefit of taking into account label interaction. Firstly, despite the natural tendency for a relatively poorer ability to classify non-substrates (as they are the minority class, and are also more prone to containing noisy data), the CC model showed more balanced single-label models that compromised slightly sensitivity to gain some specificity. This translates into a lower ΔPR measure (average deviation in precision and recall) for the CC model, indicative of less discrepancy between the ability to predict substrates and non-substrates. Secondly, two of the single-label models used other predicted labels in preference to the molecular descriptors during the CC training, leading to improved Sen to Spe balance. Thirdly, the two MRP2 single-label models within CC and BR, despite showing the same predictive accuracy performance, resulted in two very different applicability domain profiles. While MRP2-CC allowed establishing a more reliable accuracy vs STD profile, which emulates more closely the reliability profile in external data, MRP2-BR was not able to achieve this. It is hypothesized that the presence of previous label predictions allowed overcoming a systematic bias in the ensemble predictions, as this is the only aspect that changed between BR and CC. These observations consist of a proof of concept of the utility of addressing transporter overlap when modelling a QSAR, and possibly more marked effects could be obtained with a more populated matrix of instances vs transporters.

An analysis of the molecular features showed that there is some degree of overlap between transporters in terms of the molecular features responsible for substrate recognition, which supports the multi-label approach from a mechanistic standpoint.

Overall, the models revealed to be robust and of acceptable predictive performance, especially considering the complexity of trying to uncover unspecific mechanisms of substrates recognition by the ABC family members.

5. Using Multi-label Classification to Explore the Link among the Solute Carriers (SLCs) Transporter Family

5.1. Introduction

Following the previous chapter, where the ABC transporters were explored due to their important role in distribution, in this chapter the same rationale is applied to the Solute Carriers (SLCs) superfamily. These transporters play a key role in the ADME processes and are also associated with a wide range of disease states, which makes them a target of high potential in health research. However, SLCs are one of the most underexplored families of transporters, and membrane proteins in general (Cesar-Razquin et al., 2015).

As SLCs directly affect the disposition of drugs across a wide variety tissues, they are likely to have informative value if used as predictors of Vd, which prompts the QSAR modelling of their uptake profile with the end goal of helping the Vd modelling later on, as discussed in the Workflow in Chapter 3.

In a classification problem, such as the prediction of substrates for SLC-mediated transport, where multiple responses (transport by different SLCs) coexist for an individual drug, handling each transporter response individually will undeniably overlook possible interactions between them. Such interactions carry potentially important information which may facilitate the learning of patterns that characterize the problem (Gibaja and Ventura, 2015). So, instead of developing independent (classifier) models, one for each response (called label), a multi-label classifier will incorporate all responses simultaneously. Consequently, in order to uncover the potential relationships between SLC members, a multi-label approach should be taken, where the binding profiles of different transporters are modelled together (Gibaja and Ventura, 2015). As established in the Introductory Chapter 2 and evidenced in the work in Chapter 4, Classifier Chains (CC) is a particularly suited technique to address potentially correlated responses, as this technique is able to harness the information contained in any label overlap towards improving the overall modelling performance (Gibaja and Ventura, 2014). Recall that for a set of L individual SLC labels (where each label consists of transport data for a given transporter), this technique produces L classification models which communicate the learned information to each other, in a chained fashion (Gibaja and Ventura, 2014, Gibaja and Ventura, 2015). As done for the ABC family previously (Chapter 4), by comparing a CC model with an equivalent multi-

label model that handles each transporter individually (Binary Relevance, BR), one is able to effectively test whether taking into account the data overlap (i.e. transporters' substrate/non-substrate overlap) is beneficial and, from this, conclude whether there is, in fact, a meaningful correlation between transporters.

In this work substrates and non-substrates from six SLC transporters were available from the SLC uptake dataset (see section 3.1.2), namely Peptide Transporter 1 (PEPT1), Organic Cation Transporter 1 (OCT1), and four Organic Anion-Transporting Polypeptide transporters (OATP1B1, OATP1B3, OATP1A2 and OATP2B1). Given that label order has a large impact on the predictive performance of a CC model, an exhaustive exploration of all label combinations was carried out. More information on the impact of label order and how to address this have been provided in section 2.6.2. This also allows exploring the relationships among the different SLC transporters. This is the first attempt to both using multi-label classification to model SLC data and to perform an exhaustive exploration of transporter combinations.

5.2. Methods

5.2.1. Dataset and Molecular Descriptors

The SLC dataset (see Methods Chapter 3) was split into training, internal validation and test sets in a proportion of 3:1:1. As the current problem has six different labels, and each compound is allocated to a different number of labels, this partition was done by an implementation in R of a stratification procedure (Sechidis et al., 2011) that maximizes the distribution balance of all labels and both classes (substrates and non-substrates) under each label in the three subsets. A full account of the spread of data across labels in the final dataset is presented in Table 5.1. The training set was used for pre-processing of the molecular descriptors and model development, the validation set was used for any optimization tasks (feature selection optimization and tuning of model's parameters), and the test set was used exclusively for model testing (after the selection of the best model). The dataset was annotated with molecular descriptors obtained from MOE 2013 and ACD/logD suite v12.5, as explained in Chapter 3.

5.2.1. Pre-processing feature selection.

To maximize the ability to model each transporter, five different feature selection methods were carried out, using WEKA 3.6. These are three filter methods, namely Greedy Stepwise

search (GS), Genetic Algorithm (GA) and ReliefF, and two wrapper methods, namely Random Forest-Greedy Stepwise search (RF-GS) and C4.5 (C4.5) Decision Tree-Genetic Algorithm search (C4.5-GA). For full information on the used parameters and procedure of feature selection please refer to section 3.3.

Table 5.1. Distribution of labels across the training (TR), internal validation (IV) and test (TE) subsets. S and NS denote substrates and non-substrates, respectively.

Transporters (labels)		TR	IV	TE	S:NS ratio (class imbalance)
OATP1A2	S	33	11	11	2.1
	NS	16	4	4	
OATP1B1	S	58	18	19	2.4
	NS	24	7	6	
OATP1B3	S	34	11	13	2.3
	NS	15	5	6	
OATP2B1	S	29	10	8	0.8
	NS	36	14	14	
OCT1	S	93	33	33	1.6
	NS	53	18	17	
PEPT1	S	147	51	48	2.9
	NS	50	17	14	

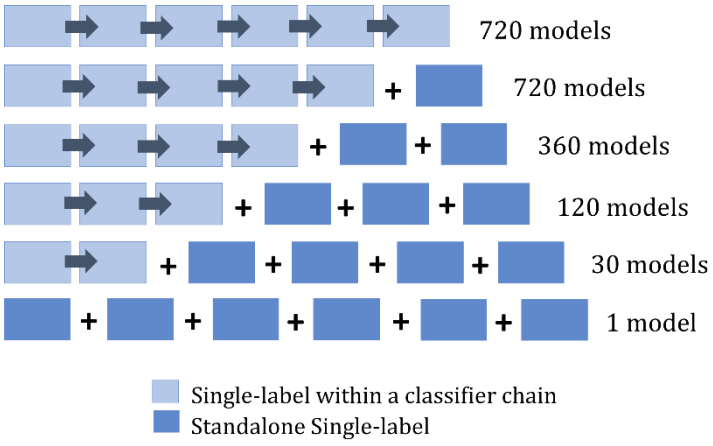
5.2.2. Multi-label QSAR modelling

As the purpose of this effort was to investigate the relationships between transporters, a CC technique was selected as the learning scheme due to its ability to account for potential transporter correlation (called label interaction in the multi-label machine learning context). If a given set of compounds has several measured response variables, e.g. interaction (or lack thereof) with several transporters (each called a label within the data mining context), there is a possibility of correlation between labels, and if correlations are indeed present, exploring them reduces the complexity of the learning task (Gibaja and Ventura, 2014).

In this work, compounds are classed as substrate or non-substrate of each of the six transporters (labels) and the SLC dataset was used to train models in a feedforward chain sequence, implemented as follows. A classifier for label #1 (the first transporter) is trained (using molecular descriptors as the model features) and feeds its prediction set (predicted class for the compounds) to the classifier for label #2 (the second transporter in the chain), which is, in turn, trained using label #1 predictions alongside molecular descriptors. The classifiers for label #2 and label #1 then feed their class predictions forward to the classifier

for label #3, and so on. Note that each label model, i.e. a model for a single transporter, within a multi-label scheme is called a single-label model. In summary, any available prior predictions will be used as a feature at any point in the chain alongside the molecular descriptors, so the classifier for label #6 will use predictions from labels #1 through #5 as additional features to complement the molecular descriptors. Predicted labels used as descriptors are generically termed “pLabel(s)”, where specific pLabels are named by prefixing the label in question with a “p” (e.g pMDR1).

To allow full exploration of all types of interactions between labels, an exhaustive search of all possible combinations of chain sequences was carried out, as shown in Scheme 5.1. This entails that all label permutations (orderings) in a 6-label chain are tested; and, for each of the possible combinations of shorter chain sizes, all possible permutations for that combination are also tested. As this problem is focused on addressing all 6 transporters, shorter chain sizes will be completed, by default, with alternative standalone single-label models, as demonstrated in Scheme 5.1. Note that standalone single-label models are originated from an alternative BR model that is built as the non-label interaction baseline comparator to the CC model. In summary, the hypotheses being tested in this study are two-fold: Is there any benefit from accounting for transporter overlap? If so, which transporters’ *substrate prediction* benefits from information from other transporters?



Scheme 5.1. Schematic representation of the exploration space of possible label (transporter) combinations. Note that all but the last line in the scheme represent different formats of the CC model of varying lengths, and the last line represents the BR alternative model.

In order to maximize predictive accuracy, prior to building the label combinations represented in Scheme 5.1, each label’s modelling conditions (classifier algorithm and feature set) were optimized and a selected single-label model was established for each

transporter. To this end, each label was modelled with each of the five available feature sets generated by the five feature selection methods, and the best classifier-feature set pairs were selected based on the performance on the validation set. This was done for three different classifiers available in WEKA: C4.5 (J48), RF (RandomForests) and boosted C4.5 trees (multiBoostAB + J48), which were tuned using 10-fold cross validation. For the C4.5 models, the pruning was tuned as per section 3.4. For the RF the number of trees was optimized (ranging between 2 and 1000 trees, with a step of 50 trees). For the boosted trees (BT), the conditions for the embedded C4.5 trees were inherited from the previously optimized C4.5 models, the number of committees (or iterations) was optimized ranging from 10 to 100, and the number of subcommittes was set to the squared root of the committee size as recommended by the author (Webb, 2000). Additionally, whenever a classifier failed to generate a good model for a certain label (G-mean < 0.7, where G-mean is defined in section 3.5), the algorithm was re-run using the feature set that previously generated the model with the highest G-mean, and a misclassification cost (optimized between 2 and 6) was applied to penalize mispredictions of the minority class – that is, the cost of misclassifying an instance of the minority class is multiplied by a number between 2 and 6, whilst the cost of misclassifying an instance of the majority class remains 1.

Two labels (PEPT1 and OATP1B3) produced models with non-acceptable performance (i.e., they had either sensitivity or specificity below 0.5, in the validation set) with any of the above classifiers. As a result, special methods were applied to them. Initially, the synthetic minority over-sampling technique (SMOTE) was applied following the re-running of the best modelling conditions up to this point, for each transporter respectively. This showed acceptable performance for OATP1B3, but not for PEPT1. To overcome this, PEPT1 was submitted to an under-over bagging (UOBag) procedure similar to a procedure in the literature (Galar et al., 2012), which led to acceptable performance. This consisted of a series of 10 runs where, in each run, the dataset was submitted to SMOTE, which added 50 (100%) instances to the minority class, followed by undersampling of 47 (32%) instances in the majority class (to reach two balanced classes), and 80% random sampling from the total resulting data. The sampled subset was then used to build a C4.5 model (using parameters inherited from prior C4.5 optimization) with an applied misclassification cost of 2 to each false positive prediction achieved during training. This generated an ensemble of ten C4.5 models that form the final UOBag model.

5.2.3. Performance evaluation and Applicability Domain

The single-label models contained in each multi-label model were assessed with Sen, Spe and G-mean. The multi-label performance measures used were F1, Hamming Loss, bACC, Sen-L, Spe-L (label-wise sensitivity and specificity). These have been calculated as descried in section 3.5.

To enrich the analysis of the models' performance, the external set was tested for the presence of activity cliffs. This allows understanding and pinpointing any shortcomings of the models under evaluation. To search for possible activity cliffs, the similarities between external (test) compounds and all training set compounds were computed. Any instance that has a different class from the majority class of the closest three training neighbors with a Tanimoto coefficient > 0.55 to the analysed compound was deemed to be an activity cliff.

In order to define the reliability of predictions output by the QSAR models, it is necessary to define their applicability domain (AD). In this study, the AD of all the single-label models used in the best model was characterized. To establish the AD, the standard deviation of an ensemble of models (STD) (Tetko et al., 2008) was used, as described in Section 3.6. This measure has been shown to correlate well with reliability of external predictions (Sushko et al., 2014, Dragos et al., 2009, Sushko et al., 2010a, Tetko et al., 2008). A small disparity in the ensemble (small STD) will (likely) equate to highly reliable predictions. However, one should keep in mind that small STD can also result from systematic bias. By computing the percentage of correct predictions (accuracy) within increasing thresholds of STD scores, one is able to establish the model's AD profile by sorting the data into higher reliability areas and lower reliability areas. As the different test sets are relatively small, during the binning of the data each STD threshold step was required to include at least 3 new instances. This is done to avoid bins composed of too few compounds, thus making each point minimally representative of the occupied STD range.

5.2.4. Statistical tests.

To objectively determine the significance of the difference in performance between the best CC model and the baseline method, the BR model, a paired Wilcoxon signed-rank test was carried out (given that all Shapiro-Wilk tests indicate non-normal distributions), according to expert recommendations (Japkowicz and Shah, 2011). The test compares the performance of a set of equivalent (between classifiers) and independent trials performed for the two single-label classifiers. This was adapted to a multi-label setting, by considering as a trial the set of multi-label predictions for each instance. This allows conserving the trial

independence conditions, and the assumption of equivalence between the respective trials of both classifiers. The statistical test was performed on the individual elements of Hamming Loss and F1 measures (for each compound in the test set), produced by the BR and CC models.

To test the value contributed by each individual transporter, a comparison of performance between CC models with and without each of the transporters was also done using the paired Wilcoxon signed-rank test.

As all analyses were performed in a large sample size, the appropriate large-sample Z approximation to the statistical test was applied to avoid misleading significant differences brought by a test score that approaches normality (Corder and Foreman, 2009).

5.2.5. Visualization of chemical space.

In order to visualise the chemical space of the SLC data and analyse the overlap of different transporters, t-SNE (Maaten and Hinton, 2008) was applied for multidimensional projection as detailed in Chapter 3. A single run of t-SNE projection was applied to the full set of compounds in the SLC multi-label dataset.

5.3. Results and Discussion

5.3.1. Multi-label model optimization and testing.

Prior to training the multi-label models, the modelling conditions for each label (transporter) were optimized. From a set of 5 possible feature selection techniques, and 3 possible classifiers (with and without misclassification cost), the best combination of features, classifier and cost was optimized for each label, based on the highest internal validation performance (best G-mean). Table 5.2 summarizes the optimal conditions achieved for each label.

As shown in Scheme 5.1, a total of 1951 multi-label models were produced (1950 CC models + 1 BR model). From these, the best model was selected based on the highest average bACC on the validation set. In this CC model, the order of labels was OCT1, OATP2B1, OATP1A2, PEPT1, OATP1B1 and finally OATP1B3. Its predictive performance on the test set, at a single-label level, is summarized in Table 5.2. The results can be compared with Table 5.3, where the test performance of the corresponding single-label

models from the BR multi-label model (non-label-interaction equivalent of CC) is summarised.

Table 5.2. Single-label test set performance of the best CC model. Predicted transporter binding labels used as features are generically presented with the respective transporter prefixed by the letter "p".

Label	Best classifier	Best feature selection	Best misclassification cost	Sen (%)	Spe (%)	G-mean (%)	Labels present as features in the models
#1: OCT1	C4.5	C4.5-GA	Equal cost	90.9	64.7	76.7	Not applicable; first label in chain
#2: OATP2B1	C4.5	RF-GS	1.3*FN	50.0	85.7	65.5	pOCT1
#3: OATP1A2	BT	GS	Equal cost	90.9	25.0	47.7	pOCT1, pOATP2B1
#4: PEPT1	C4.5 (UOBag)	GS	2*FP	81.3	57.1	**68.1**	pOCT1, pOATP2B1, pOATP1A2
#5: OATP1B1	RF	C4.5-GA	3*FP	68.4	83.3	75.5	pOCT1, pOATP2B1, pOATP1A2, pPEPT1
#6: OATP1B3	RF (SMOTE)	GA	Equal cost	92.3	66.7	**78.4**	pOCT1, pOATP2B1, pOATP1A2, pPEPT1, pOATP1B1

Table 5.3. Single-label test set performance for the BR model equivalent to the best CC model.

Label	Sen	Spe	G-mean
OCT1	90.9	64.7	76.7
OATP2B1	50.0	100.0*	**70.7**
OATP1A2	90.9	25.0	47.7
PEPT1	68.8	57.1	62.7
OATP1B1	68.4	83.3	75.5
OATP1B3	84.6	66.7	75.1

At a multi-label level (i.e. looking at the set of classifiers), the CC model showed improved performance compared to the BR model (Table 5.4), across almost all performance measures, with the exception of Spe-L. However, this can be attributed to an increase in Spe from 85.7% to 100% for the OATP2B1 label within the BR model, as the remaining

single label models have the same Spe values in Tables 5.2 and 5.3. Even though both HL and F1 are not statistically different between models ($p > 0.05$), this was expected given that both models output a very similar set of predictions, which does not allow for a differentiation at a global scale.

Table 5.4. Multi-label performance obtained on the test set. Recall that HL is to be minimized, whilst the other measures are to be maximized. Statistical testing was only carried out for the instance-based measures, as explained in the Methods section 5.2.

	measure	**BR**	**CC**	**p-value**
Label-based	bAcc	68.1	**68.7**	n.a.
	Sen-L	75.6	**79.0**	n.a.
	Spe-L	**66.1***	63.8	n.a.
Instance-based	HL	26.4	**22.8**	p = 0.164
	F1	70.7	**71.9**	p = 0.128 (S); p= 0.671 (NS)

n.a. – non applicable. *results from single-label performance measure marked in Table 5.3 (see text).

In order to understand the factors that led to the superior multi-label performance of the CC model, both models were compared at the single-label level. Tables 5.2 and 5.3 show that three out of the six labels in the multi-label CC model show no improvement over the standalone single-label models; on the other hand, PEPT1 and OATP1B3 show improvement in the CC model, which results from a marked increase in Sen accompanied by maintained Spe. The classifiers for both PEPT1-CC and OATP1B3-CC selected all available previous labels as predictors, and considering that the introduction of previous label predictions was the only differing aspect between the BR and the CC models, it can be concluded that this was the driving factor of the improved performance of PEPT1-CC and OATP1B3-CC. Moreover, it should be highlighted that all previous label predictions used in the modelling of PEPT1 refer to transporters that seemingly have no obvious structural (Schlessinger et al., 2013b) or physiological link (Cesar-Razquin et al., 2015) to it (Schlessinger et al., 2010). The only common denominator between PEPT1 and the remaining transporters is the fact that their families (SLC15, SLC22 and SLC21/SLCO) all belong to the alpha-group of the major facilitator superfamily, which entails some level of homology in their type of folding (Höglund et al., 2011). It is known that PEPT1 binding relies on very specific three-dimensional requirements of distance and chirality(Foley et al., 2010) which may be difficult to properly portray even when using three-dimensional molecular descriptors. The use of predicted labels (prediction of binding to other transporters) as

additional features may supplement the molecular descriptors used in this work to aid a better prediction of PEPT1 uptake.

The SLC22 (where OCT1 belongs) and SLCO (where the OATPs belong) families show a relatively high degree of aminoacid sequence similarity (Schlessinger et al., 2013a, Schlessinger et al., 2010). These observed links between transporters in the CC model will be later explored in more detail.

Regarding the CC model's accuracy, except for OATP1A2, which showed very low Spe, all the other labels show acceptable predictive performance despite the high degree of class imbalance for most transporter labels. It should be noted that OATP1A2 has been associated with the largest chemical space among OATPs (Tirona and Kim, 2014), which potentially makes the task of modelling its substrate recognition patterns more difficult. The sheer quantity of pattern types is perhaps larger than the instances occurring under each pattern, which makes for a much shallower input provided to the learning algorithm. This possibly justifies the poor performance obtained for this transporter. Nonetheless, the obtained multi-label performance of the CC model (Table 5.4) showed a good level of quality with a 22.8% error rate (Hamming Loss) across all compound-transporter pairs, and an F1 score above 70%. However, as usual in any QSAR effort, all transporters' models have an overall higher propensity for correctly predicting substrates than non-substrates. This makes the CC model better suited for early screening (as opposed to late-stage prediction) where a higher false positive (false substrate) rate is less damaging to the long term success of the drug development project than the opposite scenario (higher false negative rate, which leads to false hits being moved forward across the pipeline).

Regarding the source of mispredictions in the CC model, one might consider the fact that the data used in the modelling comes from different sources that employ a variety of experimental designs and analytical methods to assign a compound as a substrate or non-substrate. Different laboratories can yield large variations in transporter function and expression. Furthermore, differences in experimental conditions potentiate the variability of obtained results (Artursson et al., 2013). This generates noise, which hinders the ability of the algorithm to properly learn any patterns between structure and response. Furthermore, as pointed out by Tu et al (Tu et al., 2013), any transporter substrates capable of high passive permeability will not be detected as substrates in permeability experiments. Consequently, actual substrates will be wrongly assigned to the non-substrate class, which will hinder the learning task as similar structure patterns will be found in both the non-substrate and substrate classes. In this work, where a compound had been reported as both as substrate and non-substrate of a transporter by different literature sources, the

compound was assigned to the substrate class following the principle of minimum evidence of substrate capability (i.e. even one single positive evidence of substrate capability shows that a compound is a substrate under certain experimental conditions). There was a total of 5 observations in the test set with such conflicting literature class assignments, among which only one was mispredicted under the CC model.

5.3.2. Model Validation

The nature of the automated exhaustive search through the different classifier chain orders does not accommodate an in-process trimming of potential outliers detected through AD analysis, before a set of predictions is fed forward across any given classifier chain. Furthermore, even if such in-process control was feasible, it would likely increase the degree of overfitting as outlier removal would be guided by an AD established using the validation set. This would be, to some extent, as controversial as simply removing mispredictions from the prediction set of a label before it is fed to the following labels. As a result of this, the AD analysis established for the best model (CC) is reserved to a prospective use, where it will aid the prediction acceptance decision. The contribution of outliers fed to any of the following labels during the training of a chain will likely be picked up during AD analysis of external predictions, as decision paths constructed from such outliers will be associated with a higher level of disparity between ensemble models. As a direct result, new instances that use such decision paths will be more likely to fall outside the AD for an established acceptance threshold (using the STD method). The established AD profiles (Appendix II, Figure A2.1-6) should be used with care as the amount of data in the test set of some transporters is so small that it might not convey representative accuracy values. This can be seen in the OATP2B1 model, where the level of precision does not correlate with accuracy. Examples such as this one are extremely challenging to characterize with respect to their AD. Nonetheless, every established AD allows the selection of a subset of the data associated with a reduced risk of mispredictions (even OATP2B1 where the lowest STD threshold shows the lowest error rate).

Lastly, in order to properly gauge the value of the built models, their prediction performances have to be evaluated in light of the content in activity cliffs. Knowing that these are located in areas where the structure-response relationship is more complex and changes unpredictably (Maggiora, 2006) or they result from incorrect class assignments (Sedykh et al., 2013), one would expect that the machine learning algorithm will mispredict them. Results (Appendix II, Table A2.1) show PEPT1 data contained a considerable amount of activity cliffs, which also indicates it is a challenging label to model as discussed before.

Other than PEPT1, all other transporters were associated with a relatively low number of activity cliffs, in comparison with the number of mispredictions. This shows that activity cliffs content is not likely to have considerably hindered the learning tasks.

5.3.3. Impact of each transporter label on the global predictive performance

To assess the impact of each label on the best CC model and the BR model, each label was isolated and the performance of the remaining multilabel model was calculated using Hamming Loss for the BR model. While the BR architecture allows simulating the full classifier without each of the existing labels, this kind of analysis is not possible for the CC model. The CC model is built in an incremental manner; therefore the only way to test a label's impact on the model is by measuring the HL of the chain upon the addition of each label to the chain. Note that labels that are not in the chain are used in the model anyway, but they are used as non-linked single label components (similar to the BR model).

Figure 5.1(1) shows that, while most labels appear to have similar impact over the global performance of the BR model, removing OCT1 is associated with a larger-than-normal penalty (i.e, increase in Hamming Loss), which indicates that the respective single-label model is contributing with a high predictive performance. On the other hand PEPT1 is contributing with the largest amount of error, as removing it from the multi-label BR model leads to a marked decrease in the HL value (recall this is an error measure). Given these observations alone, PEPT1 would be expected to be one of the labels that potentially benefit the most from being in a multi-label setting that utilises label interaction (i.e. the CC scheme), while OCT1 would be expected to offer support in the modelling of other labels. Both expectations were indeed observed through the marked improvement of the predictive performance of PEPT1 in the CC model compared to the BR model (Tables 5.2 and 5.3), and through the fact that OCT1 occupies the first position in the best CC model (Table 5.2).

The analysis of the label impact on predictive performance in the CC model shows an overall decrease in Hamming Loss as the chain grows (See Figure 5.1(2)). This shows that each new label is modelled with an added level of accuracy when compared with BR single label models. Recall that the set of labels shown in each point of Figure 5.1 is completed with the remainder BR single-labels, so each newly added label is replacing its BR equivalent in the prior iteration. Taking this into account, this multi-label scheme showed to be robust to any noise across the 6-label CC model. The data point corresponding to the "no interaction chain" refers to a setting where there is no link (or chain) connecting the labels, and they are modelled independently from each other. This showed to be poorer or equivalent is performance to any stage of the construction of the CC model.

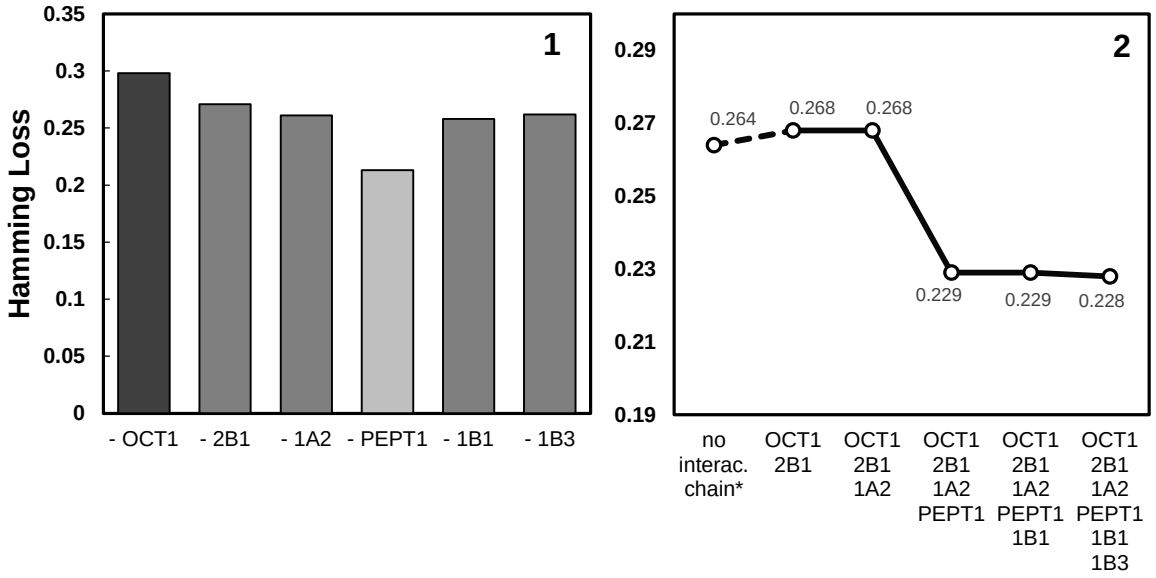


Figure 5.1. **(1)** The impact of each label over the global Hamming Loss of the BR model, computed on the test set. The impact is measured by calculating the HL of the full multi-label model upon removal of each label. Recall that HL is meant to be minimized. **(2)** The impact of replacing the single labels from the BR model with the single label components of the CC model with increasing chain length. The impact is measured by calculating the HL of the full multi-label model upon addition of a new label to the chain (rather than using the BR equivalent of the single labels). The order of labels in the chain follows that of the selected CC model. *The term “no interac. chain” refers to the scenario where there are no links between labels (i.e. the BR model). The dashed line connecting the first and second data points in the CC plot conveys the discontinuous nature between the two.

It is worth noting that the significance of previous labels is seen throughout all the models generated (in the exhaustive combinations of chain sequences) and not just in the best model discussed above. As such, an exhaustive analysis of label contribution in a multi-label modelling context showed that previous label predictions were very frequently selected as a descriptor for the modelling of any given transporter, which demonstrates that the value of using transporters as predictors among each other was not an exception found in the best achieved model. This is another evidence supporting the correlation between transporters with respect to their substrate profiles.

Looking into two-label chains where the predicted label #1 is used as the only predictor of the label number #2 can also provide useful information about how labels relate to each other. To this end the first transporter label in each possible 2-label chain was modelled using the optimal conditions (as done for all of the other CC models in this study), and its output was used as the only descriptor to model the following label in the chain using the C4.5 (decision tree) algorithm. From this exercise there are only two possible outcomes: either a one-descriptor tree is produced (with pLabel being the descriptor) or no tree is produced as pLabel is not statistically significant for the partitioning of label #2. This process

is more appropriate to test the relationships between the different transporters than running a statistical test of each correlation, given that the latter implies a symmetric (bidirectional) correlation and the former only assumes unidirectional correlations (while allowing the identification of bidirectional correlations). Given the complex nature of the problem under study it is possible that some of the relationships between transporters are in fact asymmetric (unidirectional), where transporter A offers information relevant to B, but B does not do the same for A (Gibaja and Ventura, 2014).

The summary of results in Table 5.5 (criterion B) identifies three links: 1) OATP1B1 and OATP1B3 are shown to mutually correlate to each other, as the pLabel from each transporter is selected into a decision tree to model the other transporter; 2) pOATP1A2 was selected as a predictor of OATP1B3, and 3) pOCT1 was identified as a predictor of PEPT1. Surprisingly pOATP2B1, which appears in second position in the best CC model, was not selected as a predictor for any of the transporters. However assessing labels in a two-by-two fashion is perhaps a harsh method to ascertain the significance of relationships between transporters. For example pOATP2B1 may be a predictor for a sub-group of compounds already partitioned by a molecular descriptor, rather than for all compounds. As a label receives a certain input, this can alter significantly the learned patterns, especially if different sources of input complement each other's information. As a result, the binding patterns of the OATP2B1 label, for example, might be learned very differently, which will transform what this label outputs to the remainder of the classifier chain, thus rendering it a potentially advantageous predictor of the following transporter models.

Additionally, two factors might explain the absence of other relationships present in the selected CC model from the two-label chain analyses above. Firstly, C4.5 is clearly a suboptimal learning algorithm as, in many cases, it was not the optimal training algorithm (Table 5.5). This was used here for its straightforward and transparent output, and it may have overlooked weaker correlations. Secondly, some of the correlations might not be global (and may occur in a specific region of chemical space), hence not being observed without any additional chemical information.

To assess whether there is any link between a single-label's predictive performance and its position in the CC model, the top 10 performances of a given label, at each of the six possible positions, were averaged. Figure 5.2 shows that all labels benefit, though to different extents, from being located somewhere between the second and the last position of the classifier chain as opposed to being in the first position (where no information from other labels is available). This is another evidence in support of the hypothesis of

intercorrelation between the transporters' binding profiles – this essentially means that all transporters benefit, to some degree, from previous label information.

Table 5.5. Summary of proposed links between SLC transporters, determined from four different approaches. Criterion C is a summary of the results presented in Appendix II, Table A2.2, and criterion D is derived from the results presented in Appendix II, Table A2.3.

	Criterion A	**Criterion B**	**Criterion C**	**Criterion D**
endpoint	top 5 predictor in the best model	Statistically significant predictor in a two-label chain	Statistically significant in **Obs x Obs** Chi-Square correlation	Statistically significant in **pLabel x Obs** Chi-Square correlation[#]
OCT1	n.a.	none	none	n.a.
OATP2B1	**pOCT1**	none	none	none
OATP1A2	**pOCT1**	none	**OATP1B1** **OATP1B3**	none
PEPT1	**pOCT1** **pOATP1A2** **pOATP2B1**	**pOCT1**	**OATP1B1**	**pOCT1**
OATP1B1	none	**pOATP1B3**	**OATP1A2** **OATP1B3** **PEPT1**	none
OATP1B3	**pOATP1B1**	**pOATP1B1** **pOATP1A2**	**OATP1B1** **OATP1A2**	**pOATP1B1**

Each observed endpoint was only tested with the eligible pLabel variables (i.e. the pLabels from the transporter models in lines above it, which were made available during its training). OATP1A2, for example, has two possible pLabels against which it is tested (pOCT1 and pOATP2B1) which precede it.

In agreement with other observations discussed earlier in this chapter, OATP1B3 was the transporter that most benefitted from being pushed further towards the end of the chain, showing an overall trend (except when in the second position) of increasing predictive performance as its position approaches the end of the chain. On the other hand, observations regarding OCT1 and OATP1B1 indicate that despite benefitting from being trained with information from other labels (i.e. trained later in the chain's order), these transporters show the least extent of benefit from this. This is evidenced by the smallest improvement in predictive performance from being at the top of the chain compared to being at any of the following positions. This aligns with the fact that OCT1 occupies the first position in the best multi-label model.

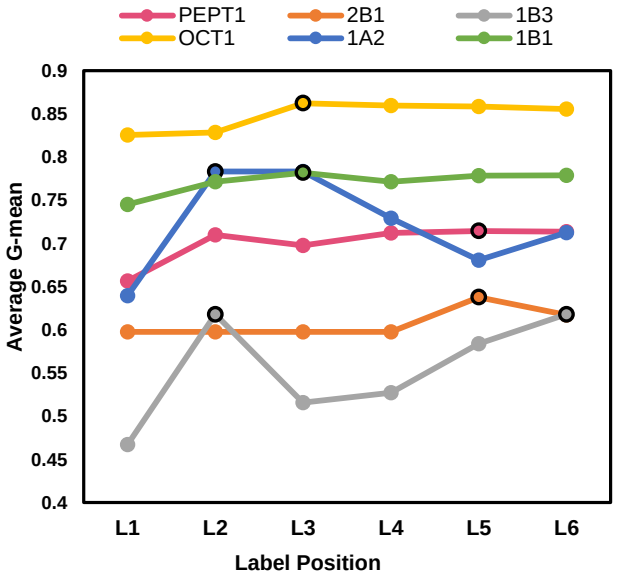


Figure 5.2. Average over the top 10 G-mean of each class label at every position in the 6-label chain. The highest G-mean points are marked with a black outline.

5.3.4. Features determining SLC binding

Looking at the relevance of features in a model can provide clues to their relationship with the modelled response. The feature importance was computed from the frequency of compounds that pass through the decision split(s) based on that feature. This measure was also corrected for node purity (i.e. ratio of correct instances in a given decision split). The latter feature importance measures are presented in Table 5.6, for BR and CC models, and the former measure is available in Appendix II, Tables A2.4 and A2.5.

As shown earlier, in the CC model all prior label predictions were selected by the transporter models in the chain (See Appendix II, Table A2.6 for the summary of molecular descriptors and Appendix II, Table A2.7 for a full list of used features). The current and following sections will discuss these findings in more detail, focusing mainly on the predicted label features. Recall that previous label predictions are represented by the name of the respective transporter prefixed with the letter “p”.

OCT1 was not applicable for modelling using previously predicted labels, as it was the first label in the chain. For the second label in the chain, namely OATP2B1, predicted OCT1 binding (pOCT1) revealed to be informative in distinguishing between OATP2B1 substrates and non-substrates, being used to sort out a third of the training set. It appears that OCT1 substrates can be both substrates and non-substrates of the OATP2B1 transporter, depending on the chemical context, whereas OCT1 non-substrates are likely substrates of OATP2B1. The single alteration of adding the pOCT1 descriptor from the BR model to the

CC model led to an attenuation of what seems to be a pronounced overfitting in the BR model (maintained Sen score, but Spe changes from 100% to 86%), as the decision splits that follow the split with a_don have a smaller combined error rate in the CC model (95% versus 84% combined node purity in the BR model).

Table 5.6. Descriptor importance for the BR model, measured in percentages of predicted and correctly predicted instances covered by each of the descriptors. For the sake of simplicity this table only shows up to the 10th most important feature, however some models used more features, as shown in Appendix II, Table A2.7. Their definitions are available in Appendix II, Table A2.6.

OCT1		**OATP2B1**		**OATP1A2**	
BR (% correct N)	CC (% correct N)	BR (% correct N)	CC (% correct N)	**BR (% correct N)**	CC (% correct N)
CASA- (96.5)		PSA (93.0)	PSA (96.9)	vsurf_EDmin2 (28.6)	vsurf_EDmin2 (26.2)
LogD7.4 (84.8)		vsurf_HB2 (75.8)	vsurf_HB2 (79.7)	Nratio (25.0)	Fu (24.3)
PM3_dipole (78.5)		PEOE_VSA_FPNEG (65.0)	PEOE_VSA_FPNEG (69.0)	NumRings6 (23.9)	NumRings6 (24.3)
a_aro (38.4)		PEOE_VSA_FHYD (51.2)	PEOE_VSA_FHYD (55.1)	Fu (22.2)	**pOCT1** (23.4)
vsurf_HB6 (30.3)		a_don (34.3)	a_don (38.2)	LogD5.5 (19.6)	a_don (23.3)
lip_violation (21.4)		AM1_E (28.1)	**pOCT1** (32.0)	SlogP_VSA1 (19.4)	SlogP_VSA1 (16.0)
vsurf_ID2 (15.8)			PM3_dipole (21.3)	vsurf_DD13 (18.0)	FASA_H (12.9)
FCASA+ (3.8)				a_don (15.3)	Nratio (11.5)
Q_VSA_PNEG (4.1)				FASA_H (10.6)	PEOE_VSA-1 (10.7)
				PEOE_VSA-1 (10.3)	LogP (9.1)
PEPT1		**OATP1B1**		**OATP1B3**	
BR (% correct N)	CC (% correct N)	BR (% correct N)	CC (% correct N)	BR (% correct N)	CC (% correct N)
AM1_HF (83.1)	AM1_HF (62.2)	FiA (34.4)	FiA (39.0)	vsurf_ID6 (28.8)	**pOATP1B1** (26.5)
ast_violation_ext (42.8)	**pOCT1** (53.2)	PEOE_VSA_NEG (33.7)	PEOE_VSA_NEG (28.1)	vsurf_ID5 (24.6)	vsurf_ID6 (25.6)
SlogP_VSA6 (39.9)	FiA (39.9)	vsurf_ID7 (20.6)	vsurf_ID7 (18.3)	vsurf_ID1 (21.1)	vsurf_ID1 (18.6)
Ro5 (30.5)	**pOATP2B1** (38.8)	vsurf_EDmin2 (18.5)	vdw_vol (17.8)	ast_violation (18.6)	vsurf_ID2 (18.2)
Fu (17.2)	Ro5 (24.2)	Q_VSA_FPPOS (16.7)	Q_VSA_FPOS (17.5)	vsurf_ID2 (18.5)	vsurf_ID5 (17.6)
FiA (17.3)	ast_violation_ext (17.8)	SMR (16.3)	Q_VSA_FPPOS (17.1)	NumRings6 (17.4)	NumRings6 (15.6)
a_nO (12.7)	**pOATP1A2** (13.8)	vdw_vol (16.2)	SMR (16.4)	vsurf_ID7 (16.8)	vsurf_R (13.6)
PSA (11.6)	LogD7.4 (13.0)	Q_VSA_FPOS (15.1)	vsurf_EDmin2 (16.2)	b_1rotN (16.1)	FRB[#] (12.4)
a_acc (10.5)	SlogP_VSA6 (11.5)	vsurf_CW2 (14.9)	glob (16.1)	AM1_Eele (14.3)	ast_violation (12.0)
a_hyd (10.2)	FiAB (9.7)	vsurf_Wp6 (14.8)	vdw_area (14.5)	Index of Refraction (14.1)	b_1rotN[#] (11.6)

For the modelling of OATP1A2, prior information on OCT1 and OATP2B1 predicted binding (pOCT1 and pOATP2B1) was made available in addition to molecular features in the CC model. However, only pOCT1 occupied a place in the top 5 most important features of the

OATP1A2 model (Table 5.6). Of course, the relationship with pOCT1 is of a complex nature in the boosted trees model, as can be seen from the two-label chain models discussed earlier (criterion B in Table 5.5) where the use of pOCT1 (as the single feature) did not yield a statistically significant partitioning of the OATP1A2 data. It is likely that pOCT1 can predict OATP1A2 substrate/non-substrate class for only a certain fraction of data.

Both pOCT1 and pOATP2B1 revealed to be highly informative in the prediction of PEPT1 binding profile (Table 5.6). Notably, as shown in Table 5.6, pOCT1 is used to correctly sort more than half of the training data across the decision tree ensemble. In the relevant fraction of compounds whose classification is affected by these two features in the decision tree model, predicted non-substrates of OCT1 and OATP2B1 are likely to be also non-substrates of PEPT1. This is evidenced by the fact that in six (out of ten) decision trees, the ensemble branches composed of “pOATP2B1 = non-substrate” and/or “pOCT1 = non-substrate” mainly lead to the non-substrates of PEPT1. In the fifth position of the chain was OATP1B1, which was modelled using the predictions output by four previous labels in the chain (pOCT1, pOATP2B1, pOATP1A2 and pPEPT1) along with molecular descriptors. Despite all four having been selected as features in the decision tree ensemble (as seen in the full set of descriptors used in each model, available in Appendix II, Table A2.7), they exhibit a low feature importance score in the RF model. As mentioned earlier, this should not be interpreted as insignificance of such features, but rather a possible role of fine separation between substrates and non-substrates.

Finally, for the modelling of the OATP1B3 single-label within the CC model, all previous label predictions were made available. Among these, pOATP1B1 was the top feature in terms of coverage of instances. This was expected considering that OATP1B1 and OATP1B3 are characterized by overlapping substrate specificity profiles (Kusuhara et al., 2013), as well as physiological cooperation (Karlgren and Bergstrom, 2016, Sharifi and Ghafourian, 2016) and 80% overlapping amino acid identity (Ho and Kim, 2014).

5.3.5. Relationships between transporters across chemical space

In this study four different approaches were used to uncover potential relationships between the six SLC transporters, and the resulting findings are summarized in Table 5.5 and Figure 5.3. Each of the analyses consists of a criterion in support, or lack thereof, of any relationship between any given pair of transporters. Some of these have been previously reported in the literature; however, some new correlations have been identified in this work. Overall, an ample variety of relationships is suggested.

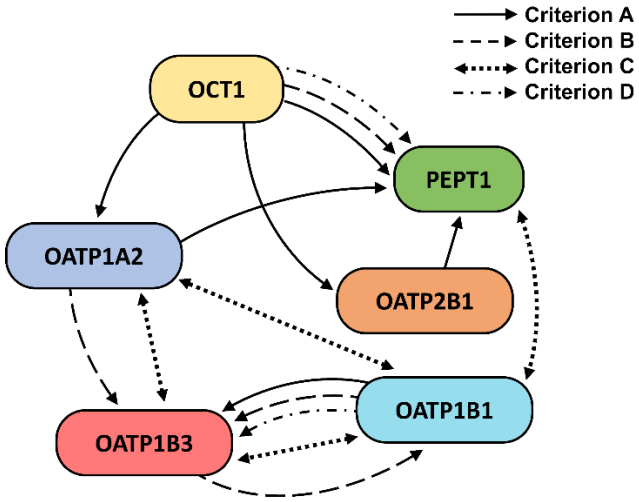


Figure 5.3. Predicted transporter relationships inferred from the four types of correlation criteria used in this work (see Table 5.5). Criteria A and B refer to predictor/predicted relationships, and criteria C and D refer to direct numerical correlations between transporters.

The identification of a predicted transporter binding as a key (top 5) predictor in the best multi-label model (criterion A in Table 5.5) is theoretically the strongest evidence of a link between two transporters and, as a result, the findings from criterion A were highlighted in Figure 5.3. This is the strongest evidence because it is the only one that simultaneously takes into account chemical context and new binding information acquired from the multi-label learning scheme. Criterion B was discussed earlier and it confirms two of the relationships proposed by criterion A; these are the significance of pOCT1 in the classification of PEPT1 and the significance of pOATP1B1 in the classification of OATP1B3 substrates/non-substrates. This criterion also suggests predictive relations between pOATP1A2 and OATP1B3, and between pOATP1B3 and OATP1B1 substrates/non-substrates.

It is worth comparing criteria C and D, as they are both statistical tests between pairs of transporters but yielded different results. The fact that some of the significant correlations found using experimental data only (criterion C) were not found using predicted data (criterion D) might result from the fact that chemical space is greatly expanded in the latter, and what might be a local correlation (for example OATP1A2-OATP1B1) might not be applicable when a broader space is considered. On the same line, correlations found in criterion C that are not present in criterion A do not necessarily mean the correlation is not important in the produced models, and it simply means that it affects a relatively small amount of compounds (hence not appearing in the top 5 important features, criterion A). As previously established, addressing specific locations of the data can be crucial for the success of a model.

In order to further understand these relationships, the distribution of various transporters' substrates across chemical space was analysed using a t-SNE projection (Figure 5.4), which shows a considerable area of overlap in the global chemical space of molecular recognition. However, Figure 5.4 also shows there are portions of space that are mainly populated by specific transporters. This will be discussed in more detail along with the label interactions found in the CC model.

A link between OATP1B1 and OATP1B3 has been demonstrated at a physiological (expression profile) (Karlgren and Bergstrom, 2016, Cesar-Razquin et al., 2015), structural (Ho and Kim, 2014), and substrate specificity (Kusuhara et al., 2013, Sharifi and Ghafourian, 2016) level, and additionally these are the major isoforms present in the liver (Tu et al., 2013). Moreover, protein quantification across a wide range of tissues shows that these transporters are both exclusively found in the liver (see Appendix II, Table A2.8). In this study the interdependence at the substrate level was also observed by using all four correlation criteria (as shown in Figure 5.3). In fact, when training the OATP1B3-CC model, the learning algorithm had access to the prediction sets of all other five transporters, but only pOATP1B1 was shown to be of considerable importance, appearing as the feature with the highest importance (Table 5.6). As for OATP1B1-CC, this label did not have access to pOATP1B3 during training, and (similarly to OATP1B3) none of the remaining available prior label features occupied a top position in the importance ranking. Therefore, as expected, Figure 5.4 shows a high degree of overlap between OATP1B3 and OATP1B1 chemical spaces.

Regarding the remaining two OATPs in this study, OATP1A2 belongs to the same family (OATP1) as OATP1B1 and OATP1B3, which entails that the three share a relatively high level of similarity, and OATP2B1 is phylogenetically more distant (Hagenbuch and Stieger, 2013). Despite the overlap of chemical space between OATP2B1 and the three members of OATP1 family, shown in Figure 5.4, pOATP2B1 did not prove to be very important in the OATP1A2-CC model (it is not one of the top 5 top predictors of OATP1A2, though it is still used as one of the predictors in the BT model), nor did it appear highly significant in OATP1B1-CC or OATP1B3-CC models. Criteria B-D in Table 5.5 also point to an overall lack of correlation between OATP1 family and OATP2B1 (in both directions). However, the argument can be made that affecting a small number of instances (as seen in the BT model for OATP1A2-CC) can, but not necessarily does, mean low importance. Depending on the specific situation, the fine decision splits in the BT model affecting a certain set of chemical compounds could hold crucial information relevant to specific portions of the chemical space (and hence not ranking highly in the feature importance measure).

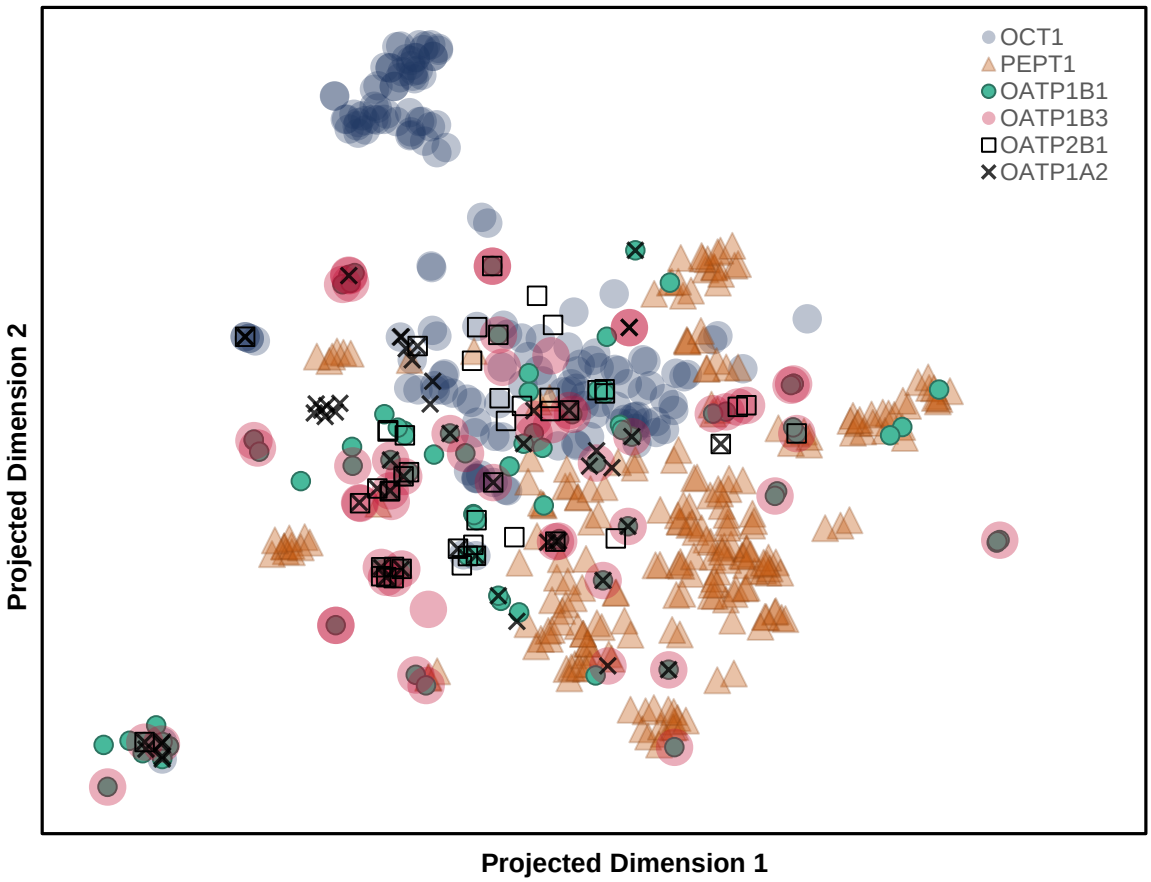


Figure 5.4. t-SNE multidimensional scaling of the Morgan Fingerprints calculated for the full SLC dataset (substrates and non-substrates). However, to allow a more straightforward visualisation, only substrates were plotted.

Focusing on OATP1A2, as expected from the relative phylogenetic proximity discussed earlier, this showed various accounts of substrate specificity correlation to OATP1B3 and OATP1B1 (i.e., in Table 5.5, OATP1B1, criterion C; OATP1B3, criterion B and C).

The most visible, and impactful, case of label interaction was seen in the PEPT1-CC model. In this model, the largest number of previous label descriptors is present in the top positions of label importance, with pOCT1 and pOATP2B1 being in the top 5, and pOATP1A2 in the 7th place. A separate t-SNE multidimensional scaling of the Morgan fingerprints (Figure 5.5) calculated for substrates and non-substrates of PEPT1, OCT1, OATP2B1 and OATP1A2 was performed to show the substrates of these few transporters more clearly than that depicted in Figure 5.4. Figure 5.5 shows that OCT1 substrate space covers a region of PEPT1's chemical space, and OATP2B1 and OATP1A2 substrates cover a different (though partly overlapping) region of PEPT1's chemical space. This is indicative of cooperation rather than redundancy of these predicted labels in the identification of PEPT1

substrates, which might be one of the reasons behind the simultaneous presence (as high importance features) of these transporters in the PEPT1 model. One evidence that supports this hypothesis is the fact that from a total of 116 rules that make up the PEPT1-CC model, 32 rules have both types of input, 37 rules have only pOCT1, and 29 rules have only pOATP2B1 and/or pOATP1A2 inputs.

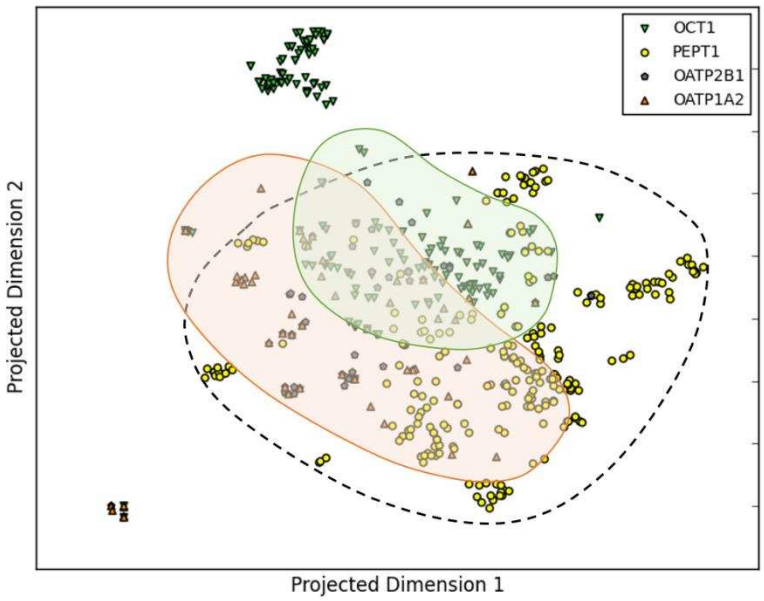


Figure 5.5. t-SNE multidimensional scaling of the chemical space occupied by PEPT1 and the prior label features present in top positions of the PEPT1-CC model. t-SNE was applied to Morgan fingerprints folded over 1024 bits. The green area corresponds to the OCT1 occupancy in chemical space, and the orange area corresponds to OATP2B1/OATP1A2. The dotted line corresponds to PEPT1's region.

The link between OCT1 and PEPT1 is further supported by the fact that pOCT1 is both picked as a descriptor in a single-descriptor tree model of PEPT1, and that these two transporters' substrate/non-substrate class show significant correlation (see Table 5.6, criterion B and C, respectively). As explained earlier, PEPT1 has not shown any type of correlation or link to any of the transporters studied here in terms of expression profile or structure similarity (which were the only large-scale analysis of correlation carried thus far, as mentioned earlier), and yet it shows to benefit the most from information regarding other transporters' binding.

The predicted binding profile of OCT1 (pOCT1) appears to hold useful information regarding other transporters, namely OATP1A2 and PEPT1, as well as OATP2B1. In the special case of OATP2B1, despite the low feature importance of pOCT1, it has been shown to clearly benefit the prediction. To be more specific, the introduction of the pOCT1 descriptor in the OATP2B1-CC decision tree only affected 21 (a third of) instances, however this factor led

to a decrease in classification imprecisions (impurity at the final nodes) from 3 to 1 (see Figure 5.6 and 5.7). The new split created by this feature led to a new, higher quality decision path, and as the trees of both BR and CC models are the same up to this point, this makes pOCT1 arguably one of the most important features to address OATP2B1 transport. Table 5.6 shows various occurrences of OCT1 (either observed or predicted) as being correlated to various transporters listed above.

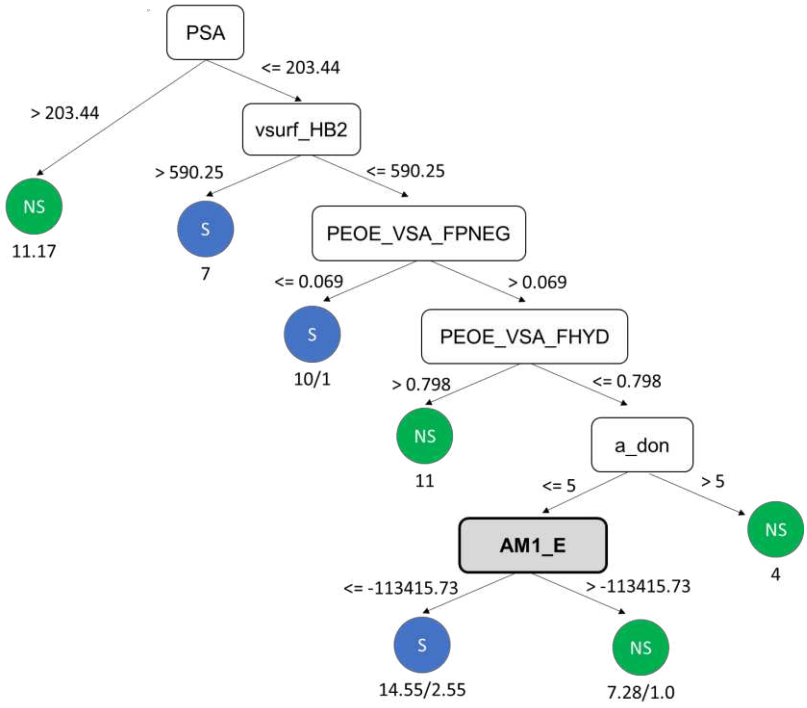


Figure 5.6. Modelling OATP2B1 substrates and non-substrates without information from other transporter labels. Compare this to Figure 5.7 where, upon introduction of prior label information, the decision tree is maintained exactly the same and the last node (in grey) is replaced by pOCT1, this allowing further splitting.

According to data reported in the human protein atlas (www.proteinatlas.org), OCT1 is widely spread throughout the body, being the most ubiquitous transporter of this study (see Appendix II, Table A2.8), and OATP1A2 (Roth et al., 2012) and OATP2B1 (Tamai and Nakanishi, 2013, Yarim and Koksal, 2010) are also expressed in a variety of tissues as depicted in Appendix II, Table A2.8, being co-expressed with OCT1 for several of their locations (the breast is the only exception). As for PEPT1, it is highly expressed in the intestine (Tashima, 2015) and the gallbladder (Appendix II, Table A2.8). Despite being only co-expressed with OCT1, its QSAR model selected OATP1A2 and OATPB1 as predictors as well, which points to an interaction between transporters that goes beyond simple co-expression in the same tissues. Furthermore, despite OCT1 being one of the main hepatic

uptake transporters of drugs, its exact physiological role beyond this is still speculative.(Liang et al., 2015)

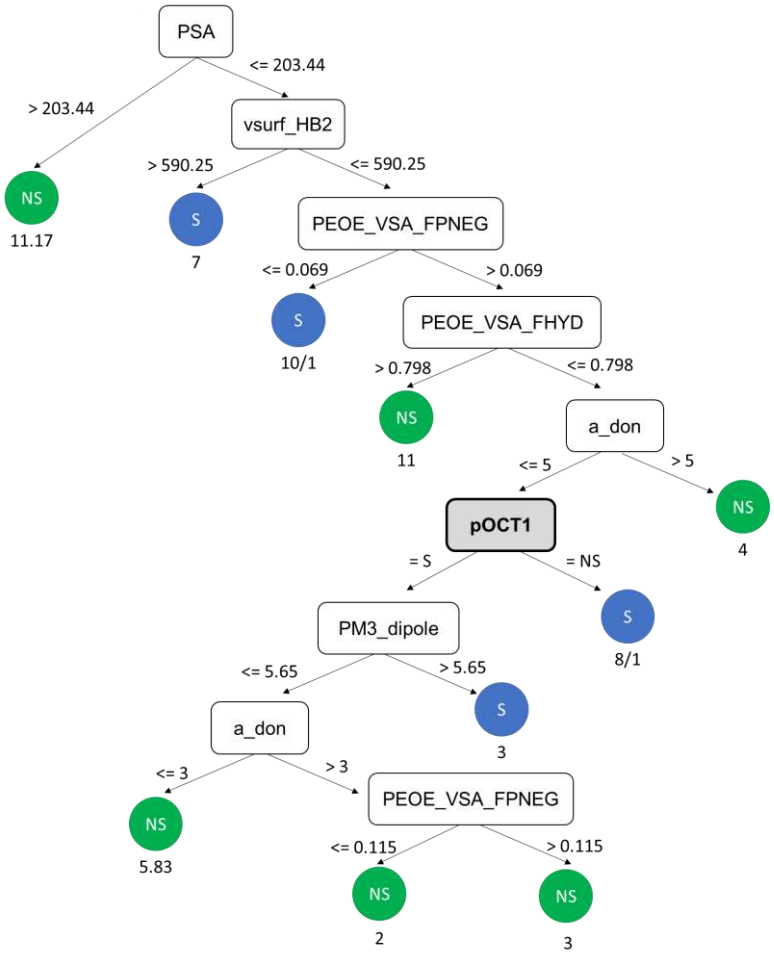


Figure 5.7. Modelling OATP2B1 substrates and non-substrates with information from other transporter labels. Compare this to Figure 5.6 where the introduction of prior label information (pOCT1) allowed further splitting and improved class separation.

Driven by the expectation that the activity of any given SLC transporter is likely to affect, or be correlated to, the activity of other members of this transporter superfamily (Cesar-Razquin et al., 2015), an extensive exploration of relationships between co-expression profiles (at RNA level) was only able to identify statistically significant correlations between OCT1+OATP1B3, OATP1B1+OATP1B3, OCT1+OATP1B1 and OATP1A2+OATP2B1 (considering only the six transporters covered in this study) (Cesar-Razquin et al., 2015). It has been speculated that co-expressed proteins across tissues and conditions (like the examples above) are functionally dependent (Cesar-Razquin et al., 2015). While evidence of relationships between some of these transporters' substrate spaces was found, OATP1A2 and OATP2B1 do not show to correlate in terms of substrate profile. In addition,

several other substrate specificity correlations were uncovered, which have not been reported to date. Some of the new proposed correlations are not apparent, for example OCT1+OATP1A2 or PEPT1+OCT1, as the identified pairs transport substrates of different chemical nature. However, it is apparent that the propensity to be transported by one transporter can predict the substrate/non-substrate class of the other.

Overall, the most important practical implications from the observations gathered in this work regarding the connection among the SLC family, encapsulated in Figure 5.3, are two-fold: firstly, if a compound has been tested for uptake against, for example, OCT1, OATP2B1 and OATP1A2, there is the possibility of predicting the binding to PEPT1 more reliably than just doing so from chemical data alone. For transporters like OATP1B1 or 1B3, this has direct relevance towards managing the risk for certain outcomes such as liver injury. Secondly, the hypothesis of a more complex relationship between transporters may prompt experimental exploration and lead to new discoveries of physiological drivers of drug disposition, which may be useful for aims such as new tissue-targeting approaches or better management of pharmacodynamics (activity and/or toxicity).

5.4. Conclusions

As presented in this chapter, SLC transporters are an appealing target of research as they are both underexplored and have a high potential in drug discovery and development applications. These proteins are widely associated with disease states and are also some of the main controllers of the ADME processes. At the same time the (Q)SAR studies on SLC binding, available up to this point, are both few and limited in their chemical space (i.e., performed mostly on series of analogous compounds). In this work a QSAR model of six different SLCs, namely OCT1, PEPT1, OATP1A2, OATP1B1, OATP1B3 and OATP2B1 was developed using a chemically diverse dataset. As there is some degree of overlap between binding profiles of the different SLC transporters, this study aimed to address the potential correlation between them in the QSAR modelling by using a multi-label classification technique called Classifier Chain (CC) that utilizes possible label correlations to aid the learning algorithm. To explore all potential interaction between transporters, all possible chain arrangements (including chains of smaller sizes and various orders) were built.

This study reports several pieces of evidence in favour of a variety of relationships between the modelled SLC transporters. From the exhaustive exploration of a total of 1950 possible CC models, the best CC model showed an overall good predictive performance across all

transporters, with the exception of OATP1A2 model's low accuracy for non-substrate identification (specificity). Additionally, it showed improved performance when compared to Binary Relevance (BR) model that assumes no label interaction. In some cases, the prior transporter predictions had a very central role in the classifiers, as they were used to classify a large amount of compounds into substrates and non-substrates. However, in some cases (e.g. the OATP1B1-CC single-label model), these prior transporter features were involved in the classification of only a small portion of compounds. Furthermore, by analysing the ten best models at each position in the chain, for each of the six transporters, results showed that every transporter benefits from being trained with information from other transporters.

The correlations uncovered by the presence of previous transporter features in each of the trained single-label models was further explored by analysing the chemical space overlap between each transporter and its prior transporter predictors. Additionally, these findings were complemented by statistical testing of transporter correlations (using both predicted and observed compound profiles), as well as by using each transporter as the single feature in a simple decision tree model. This confirmed the already identified (or proposed) transporter correlations, like OATP1B1+OATP1B3, and uncovered new potential correlations in terms of the relations between substrate space of these transporters, including relations between different OATPs and PEPT1, which belong to different protein families.

Current knowledge on the links between SLCs is based on structure similarity or expression profile correlations. The results shown here add to this knowledge and propose that SLCs might be correlated in terms of substrate specificity, which is not covered by structure similarity or expression profile correlations.

6. The Impact of Membrane Transporters and Phospholipidosis in Modelling Volume of Distribution

6.1. Introduction

As established in section 3.8 (Project Workflow), modelling volume of distribution (Vd) is the main end goal of this thesis. After modelling some of the main transporters that drive distribution - described in Chapters 4 and 5 - the output produced by both modelling efforts will now be incorporated (as input) in the construction of a Vd model.

Recall that Vd is a measure of drug distribution, which expresses the theoretical volume in which a given dose of drug appears to be distributed, based on the observed plasma concentration (Holford and Yim, 2016), For instance, if 600 mg of a drug are administered intravenously and the resulting plasma concentration is 6 mg/L, this means that the drug appears to be diluted in 100L - this volume, however, surpasses the physiological limit. As a result, the measured Vd typically does not represent an actual physiological volume, but rather the extent of binding to any physiological structures and partition into tissue compartments.

Among the different variants of Vd that can be determined, Vd at steady state (Vss) is the most reliable (Smith et al., 2015), as drug input equals the rate of output. As a result, Vss is the net result of intracellular space access and the extent of binding to various tissue and plasma components, when all these processes reach equilibrium (del Amo et al., 2013).

As can be anticipated by this property, distribution is a key determinant of the drug's ability to reach its target tissue in required concentrations, hence determining the administered dose. Additionally, distribution is also decisive in determining drugability, as it might provide clues to the ability of a drug to reach target tissues and/or off-target tissues (which might lead to toxicity issues). In fact, Vd has been correlated to the likelihood for toxicity (Sutherland et al., 2012).

An example of this is compound GEN-203, which exhibits high levels of distribution into tissues and a high Vd, likely attributed to intracellular accumulation. This phenomenon was rationalised as the mechanism by which GEN-203 elicited severe liver and bone marrow toxicity (Hop, 2015). In scenarios such as this one, the early prediction of Vd is very useful

for flagging any potentially toxic candidates, and could potentially help preventing attrition cases not only attributed to direct PK issues, but also toxicity issues.

During the drug development process, human Vd estimates are usually obtained from in vivo animal studies, where some form of animal-to-human scaling is performed (Louis and Agrawal, 2014), but relying on animals to extrapolate human PK has been demonstrated to be complex as well as unreliable (Tsaioun et al., 2016) – as discussed in section 1.9. Alternatively, estimates obtained from in vitro tissue-binding assays are also used (Zhang et al., 2012, Yanni, 2015, Tsaioun et al., 2016). However both these methods present the major drawback of requiring the synthesis of all candidates, as well as being expensive and time consuming. This renders such approaches prohibitive as high throughput screening methods. Quantitative-Structure Activity Relationship (QSAR) is an alternative option amenable to high throughput screening of human Vd in very early stages of drug development, as QSAR modelling just relies on chemical characteristics of compounds to infer the Vd.

As established in the Introduction Chapter 1, physicochemical features such as electrostatic and lipophilic profile (Waters and Lombardo, 2010) of the molecule can determine the drug distribution to a large extent, but these are only responsible for the unspecific components of the distribution phenomena, such as passive membrane permeation. Other more specific phenomena also modulate distribution, namely transporter-mediated efflux and influx, and drug-induced phospholipidosis, with both having several documented examples of their direct effect on Vd (Dantzig et al., 2004, Smith et al., 2015, Funk and Krise, 2013). This prompted the incorporation of drug transport as well as phospholipidosis as physiological features, alongside physicochemical features, in QSAR modeling for the prediction of Vd.

For this, the 10 ABC and SLC transporters, previously modelled in Chapters 4 and 5, were used in conjunction with phospholipidosis. Combining different physiological processes is desirable given that, for example, different transporters are known to work in concert to control the distribution of compounds in polarized cells present in different tissues such as the lung, intestine, liver and kidney (Dantzig et al., 2004).

With the exception of tissue partition coefficients, used by Freitas et al (Freitas et al., 2015) as predictors of Vss, no other specific physiological phenomena such as transporter uptake or phospholipidosis have ever been used as input variable in the modelling of Vss, which is the main innovation brought by this work. There has been, however, a work (Hanumegowda et al., 2010) where the authors used Vss as a predictor of phospholipidosis which, in a sense, validates the decision to use the latter as a predictor of the former. However, while from a computational standpoint this is a valid premise (as the two are indeed correlated

with each other), conceptually Vd is the net effect of all parallel processes of binding to and permeation across different tissues and biological structures, in which phospholipidosis is included. The contribution from phospholipidosis necessarily affects the volume of distribution (even if the effect is small), however the volume of distribution has no necessary implication over phospholipidosis.

Additionally, other unspecific biological binding features (i.e. serum albumin binding and membrane binding) (Sui et al., 2009, Hollósy et al., 2006) have also been used as predictors of Vss in computational models.

This chapter explores, for the first time, the potential value, from a data mining perspective, of using transporter and drug-induced phospholipidosis data in the prediction of human Vss. Additionally, motivated by the shortage of experimental data available to annotate the volume of distribution dataset, this study will also explore the feasibility of complementing experimental data with predicted data across the different physiological predictors used in the modelling process. In order to validate the results obtained, extensive comparisons against the literature through two benchmark external test sets used by in two other works (Gombar and Hall, 2013, Lombardo and Jing, 2016) was also performed.

6.2. Methods

6.2.1. Volume of Distribution (Vd) Dataset and Descriptors

The Volume of Distribution Dataset (Section 3.1.3) was used in the QSAR modelling. As a point of reference, logVss data can be visualised in an ordered bar chart, Figure 6.1. The data retrieval, molecular descriptors (ACD and MOE descriptors) and physiological descriptors (drug-induced phospholipidosis (PL), and ABC and SLC transporters) annotation on the Vd dataset was done as described in Section 3.1.3 and section 3.2 (note that, for the current chapter, the plasma protein binding descriptor present in the dataset, described in Section 3.1.3, was not used). Recall that, since the data that composes the physiological descriptors was obtained experimentally, the missing data in these variables was completed with predictions output by QSAR models of ABC and SLC transporters (built in the two previous chapters). The same was done for PL, where predictions were obtained from a QSAR model that was trained from molecular descriptors. This process is summarized in Figure 6.2, and the full details on data sources and annotation procedures are provided in section 3.1.3. The final dataset used for the modelling consisted of log-transformed Vss (logVss), 304 molecular descriptors (MDs) and 11 physiological

descriptors (PDs), namely MDR1, BCRP, MRP1, MRP2, OATP1A2, OATP1B1, OATP1B3, OATP2B1, OCT1, PEPT1 and PL. Prediction-completed features are prefixed with “p”. Tentatively, PL was used in the modelling as containing exclusively experimental data (ePL) or containing experimental data completed with prediction (pPL). Molecular descriptors were calculated as explained in section 3.2, and the MOE 2013 version was used for this.

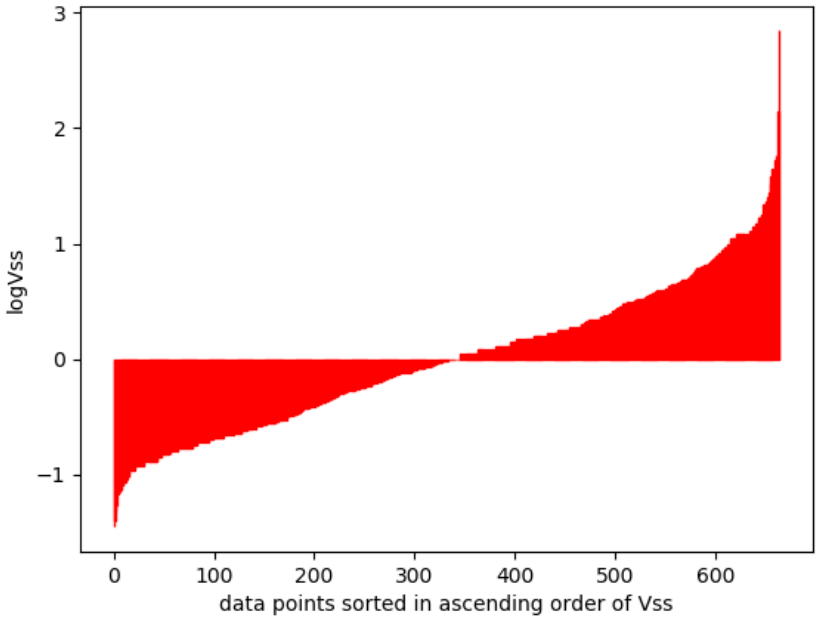


Figure 6.1. Data points ordered by ascending logVss.

6.2.2. QSAR Model Development

To create the QSAR models for the prediction of Vss, the data was split into 60% for training (N=398), 20% for testing (N=133) various modelling conditions (enumerated in Table 6.1) and for selecting the candidate models, and the remaining 20% (N=134) was set aside exclusively for final testing of the two best candidate models. All pre-processing and model training was carried using WEKA version 3.8 (Hall et al., 2009).

Two different regression methods – Random Forest (RF) and Boosted Regression Trees (BRT) – were tested using, respectively, the RandomForest function or the AdditiveRegression wrapped around the RandomTree learner, all implemented in WEKA 3.8. For the tuning of the algorithm parameters in both cases, the optimal parameter values were selected based on the lowest Mean Absolute Error (MAE) in an internal 10-fold cross validation using the training set. For the RF model, prior to modelling, the number of trees was optimized in a range between 100 and 1000 with increments of 100. As for BT, considering this was carried by wrapping a boosting algorithm around a random tree

algorithm, variable subsets of features were used to build the committee of models. The number of randomly sampled features was set to 9 (the same value used by the RF algorithm for the current dataset size), the number of iterations was optimized between 100 and 1000 in increments of 100, and shrinkage (applied to weight update) was optimized between 0.05 and 1 in increments of 0.05. All other parameters in both algorithms were used as default in WEKA.

The regression algorithms were tried in conjunction with two different correlation-based feature selection (CFS) methods, different variations of feature sets and different types of PL data. All combinations were tested to find the optimal combination of regression method, feature selection method, feature type, and PL data content, as well as to study the impact of some variables of interest (namely, feature type and the nature of the PL feature). The different variants under each experiment variable are summarized in Table 6.1.

Feature selection was tentatively performed on the whole feature set (with both PD and MD features), on PDs only or on MDs only. Alternatively, merging the feature sets obtained from separate feature selection procedures performed separately on MDs and on PDs was also attempted. This was done to cover the possibility of the feature selection methods being overwhelmed by the much larger number of MDs, which would bias the selection towards picking them over PDs.

Table 6.1. Optimized modelling parameters.

Experimental conditions to optimize	Variations tested
Feature type (provided to feature selection method)	All features physiological descriptors only molecular descriptors only physiological + molecular descriptors (selected separately)
Feature Selection Method	Genetic Search (GA) Greedy Search (GS) No Feature Selection
Regression Method	Random Forest (RF) Boosted Trees (BT)
Nature of the PL feature	Experimental PL (ePL) Experimental PL completed with predicted class probabilities (pPL)

Regarding model evaluation, the systematic comparison between different models was done using the validation set, as the test set is exclusively used for final model testing. The measures used for assessing the predictive performance are the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), coefficient of determination (R^2) and the

Geometric Mean Fold Error (GMFE), calculated as defined in the section 3.5.2. Additionally, the percentage of data within 2- and 3- fold error (FE) thresholds was calculated for the two best final models.

To analyse the role of PDs in the modelling of Vss, the content of PDs found in the best obtained model (with respect to the internal validation set) was assessed in detail. To do so, the types and number of PD combinations encountered was analysed, and the Vss coverage that they offer. In addition, this study also evaluated PDs with respect to their observed feature importance, which corresponds to the sum of the correctly modelled training compounds which depend on a given PD for their prediction (Freitas, 2013).

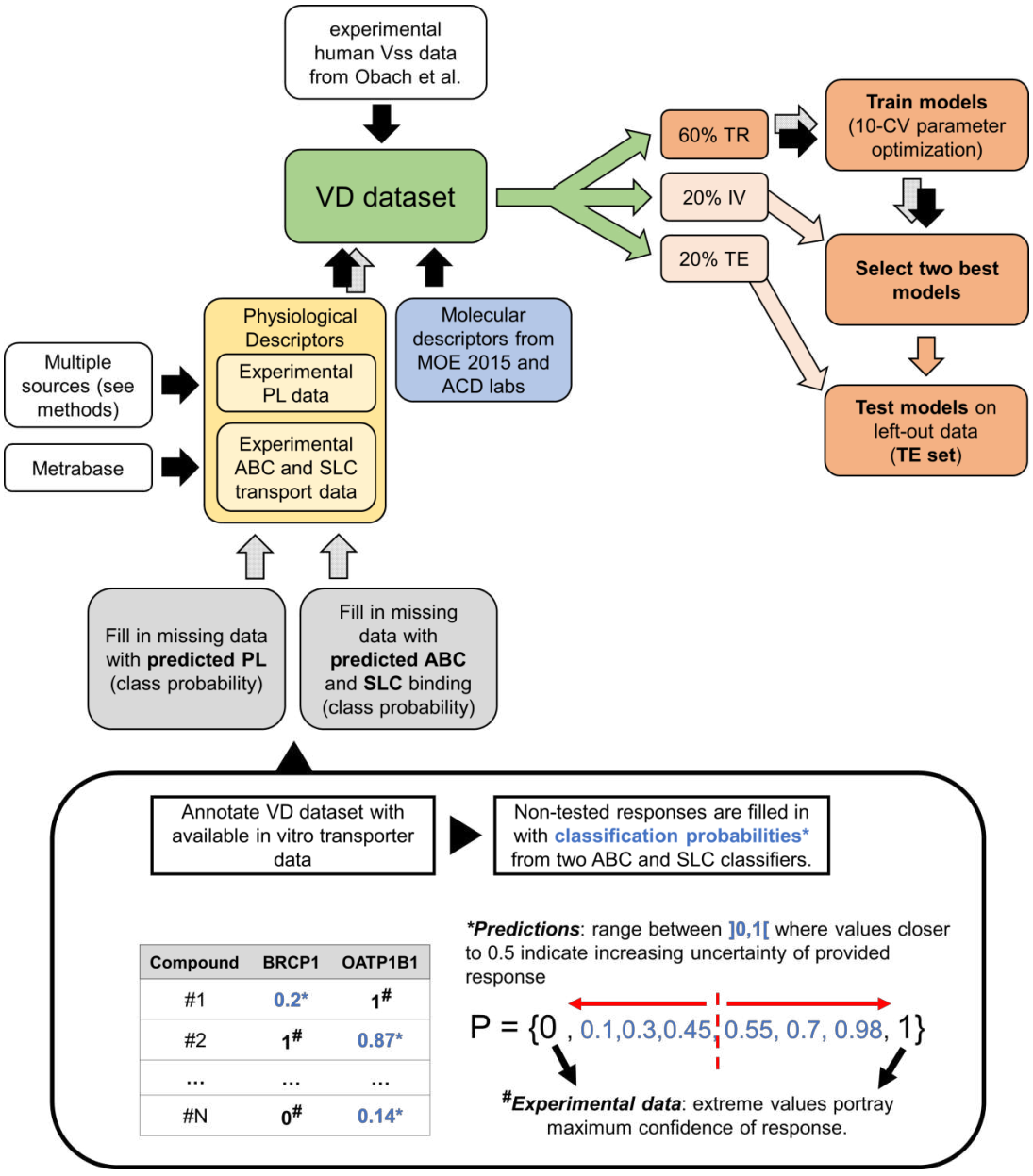


Figure 6.2. Modelling workflow.

6.2.1. Benchmark Comparison with Previous Vss Models

In order to further evaluate the best models in this work, and especially to determine the value of adding physiological information to the regularly used molecular descriptors, the best model obtained (with respect to the test set) was compared against the models reported in two other previous works by testing all models on two fixed test sets, used respectively by Gombar et al (Gombar and Hall, 2013) and Lombardo et al (Lombardo and Jing, 2016). Gombar et al trained their models on the same data in this work's dataset (Obach dataset), while Lombardo et al used an extended Vss dataset, gathered recently.

This benchmark comparison was extended by re-training this work's best model with the Lombardo Vss dataset, as well as retraining Lombardo et al's best model with the addition of PDs. For the first retraining, a new random forest was developed following a re-run of the GA feature selection step and optimization of the number of trees, since the previously used modelling parameters were optimized for a considerably smaller (Obach et al) dataset. For the second retraining, the modelling conditions optimized by Lombardo et al were directly used (since both the learner and the dataset are the same, no tuning is necessary); this consisted of a random forest of 500 trees, 33 selected molecular descriptors, 11 sampled descriptors per split, and a minimum tree node size of 10.

Note that, upon cleaning this extended dataset, 10 pairs composed of an entry with a mixture of optical isomers and another with one of its optical isomers were identified. This conflict was solved by keeping only the isomer, thus avoiding overlapping instances. Additionally, three synonyms were found and merged by averaging their Vss values.

6.2.2. Applicability Domain and Data Visualization

In order to define the feasibility of the best model developed in this work, its Applicability Domain (AD) was characterized using the STD method (Tetko et al., 2008), as described in Section 3.6. To visualize the data, t-SNE multidimensional scaling was used following section 3.7.

6.3. Results and Discussion

6.3.1. The Impact of Different Types of Input Data in Modelling Vd

As explained in Section 6.2, four sets of modelling parameters (listed in Table 6.1) were explored and optimised in this work: the feature type, feature selection method, regression method, and the nature of the PL variable. Following tuning of each algorithm's parameters, the resulting best models for each combination of the modelling parameters were tested on the validation set and the predictive accuracies were summarized in Tables 6.2 and 6.3.

When varying only the type of features used, in the great majority of cases (for 9 out of 12 modelling blocks in Tables 6.2 and 6.3, i.e. 12 different regression/feature selection combinations with either ePL or pPL) physiological information accompanied by chemical information produced the best model – comparing MAE values within each regression-feature selection block in both Tables 6.2 and 6.3, where blocks are delimited by a thicker border line. A detailed list of features used is provided in Appendix III, Tables A3.1-8. Additionally, in 7 out of 8 cases in Tables 6.2 and 6.3, models using both sets of separately selected PDs and MDs yielded lower (better) MAE values than models using just selected MDs or just selected PDs. Both these observations demonstrate that features carrying information about the physiological processes that drive distribution seem to be necessary in order to improve the modelling of Vss. Even though molecular descriptors used in isolation have not led to a markedly lower predictive performance of Vss, this underperformance is observed systematically across the various modelling conditions, which means that molecular descriptors alone are a suboptimal input for modeling Vss. Similarly, using physiological features alone was also not sufficient as the only source of input. This outcome is expected given that Vss has a large unspecific component that does not depend on transporter-mediated efflux or lysosome entrapment.

The two best models (8a and 16a) were selected for final evaluation in the test set, based on the lowest MAE on the validation set. Firstly, note that both models used pPL (rather than ePL), which follows the overall superiority of the pPL-derived models compared to models using just experimental PL data (compare Tables 6.2 and 6.3), as will be discussed in the next section. In addition, both models used a variety of physiological features. In the case of model 16a, this model was trained with features from a previous GA feature selection step. Even though there are many more features of chemical nature, various physiological features were still selected into the model (pPEPT1, pMRP2, pPL, pOCT1, pMRP1, pOATP1B1), alongside MDs (listed in Appendix III, Table A3.2). Considering that the feature selection method minimizes inter-descriptor correlation, the presence of both feature types, PDs and MDs, shows that these physiological descriptors offer additional

information to the one carried by molecular descriptors. As for model 8a, this was trained with a descriptor set that resulted from two separate feature selection runs applied to MDs and PDs. As with model 16a, various PDs were present in the model (pPEPT1, pBCRP1, pPL, pMRP2, pMRP1, pOATP1B1), alongside different MDs (listed in Appendix III, Table A3.6).

Table 6.2. Predictive accuracy on the validation set, using ePL. The number of compounds in the training and validatiom sets were 398 and 133 respectively. The two best models (selected for further analysis) from both Table 6.2 and 6.3 are highlighted in bold. Regarding the feature content available in pre-processing, "all feature" corresponds to 315 features being made available, "MDs" corresponds to 304 features, "PDs" corresponds to 11 features, and "FS-MDs + FS-PDs" corresponds to separate feature selection procedures performed on 304 MDs and 11 PDs. FS: feature selection. *both models resulted from the same input feature set, hence same performance.

Regression Method	Model	Feature Selection	Feature content available in pre-processing	Feature types present in the model	R^2	RMSE	MAE	GMFE
Random Forest	1*	GS	all features	MDs	0.431	0.4587	0.3385	2.18
	2*	GS	MDs	MDs	0.431	0.4587	0.3385	2.18
	3	GS	PDs	PDs	0.084	0.6162	0.4692	2.95
	4	GS	FS-MDs + FS-PDs	MDs, PDs	0.433	0.4577	0.3357	2.17
	5	GA	all features	MDs, PDs	0.465	0.447	0.317	2.07
	6	GA	MDs	MDs	0.444	0.454	0.318	2.08
	7	GA	PDs	PDs	0.084	0.616	0.469	2.95
	8	GA	FS-MDs + FS-PDs	MDs, PDs	0.474	0.442	**0.307**	2.03
	9	None	all features	MDs, PDs	0.473	0.447	0.318	2.08
	10	None	PDs	PDs	0.271	0.522	0.380	2.40
	11	None	MDs	MDs	0.438	0.4594	0.3266	2.12
Boosted Trees	12	GS	all features	MDs	0.194	0.581	0.441	2.76
	13	GS	MDs	MDs	0.194	0.581	0.441	2.76
	14	GS	PDs	PDs	0.011	0.794	0.599	3.98
	15	GS	FS-MDs + FS-PDs	MDs, PDs	0.371	0.487	0.360	2.29
	16	GA	all features	MDs, PDs	0.461	0.447	0.322	2.10
	17	GA	MDs	MDs	0.458	0.449	0.318	2.08
	18	GA	PDs	PDs	0.011	0.794	0.599	3.98
	19	GA	FS-MDs + FS-PDs	MDs, PDs	0.447	0.453	0.321	2.09
	20	None	all features	MDs, PDs	0.458	0.451	0.319	2.09
	21	None	PDs	PDs	0.212	0.566	0.413	2.59
	22	None	MDs	MDs	0.466	0.4468	0.3138	2.06

Table 6.3. Predictive accuracy on the validation set, using pPL. The two best models (considering both Table 6.2 and 6.3) are highlighted in boldface. The selection of these two models was based on the lowest MAE among all available models. The number of compounds in the training and validation sets were 398 and 133 respectively. *both models resulted from the same input feature set, hence same performance.

Regression method	Model	Feature Selection	Feature content provided in pre-processing	Feature types present in the model	R^2	RMSE	MAE	GMFE
Random Forest	1a*	GS	all features	MDs	0.431	0.4587	0.3385	2.18
	2*	GS	MDs	MDs	0.431	0.4587	0.3385	2.18
	3a	GS	PDs	PDs	0.160	0.5817	0.424	2.66
	4a	GS	FS-MDs + FS-PDs	MDs, PDs	0.447	0.4519	0.3273	2.13
	5a	GA	all features	MDs, PDs	0.469	0.445	0.308	2.03
	6	GA	MDs	MDs	0.444	0.454	0.318	2.08
	7a	GA	PDs	PDs	0.160	0.582	0.424	2.66
	8a	**GA**	**FS-MDs + FS-PDs**	MDs, PDs	0.474	0.442	**0.306**	2.02
	9a	None	all features	MDs, PDs	0.453	0.455	0.322	2.10
	10a	None	PDs	PDs	0.267	0.523	0.371	2.35
	11	None	MDs	MDs	0.438	0.459	0.326	2.12
Boosted Trees	12a	GS	all features	MDs	0.197	0.581	0.441	2.76
	13	GS	MDs	MDs	0.194	0.581	0.441	2.76
	14a	GS	PDs	PDs	0.053	0.769	0.588	3.87
	15a	GS	FS-MDs + FS-PDs	MDs, PDs	0.348	0.498	0.370	2.34
	16a	**GA**	**all features**	MDs, PDs	0.511	0.425	**0.304**	2.01
	17	GA	MDs	MDs	0.458	0.449	0.318	2.08
	18a	GA	PDs	PDs	0.053	0.769	0.588	3.87
	19a	GA	FS-MDs + FS-PDs	MDs, PDs	0.448	0.453	0.316	2.07
	20a	None	all features	MDs, PDs	0.482	0.441	0.316	2.07
	21a	None	PDs	PDs	0.216	0.584	0.412	2.58
	22	None	MDs	MDs	0.466	0.4468	0.3138	2.06

To compare these two models and establish which one is expected to perform better in unseen data, their performance on the test set was analysed (Table 6.4). The coefficient of determination (R^2) is larger and 5 out of 6 error measures are better for model 8a, which makes this the overall best model. Model 16a was only better in terms of the percentage of compounds with predicted Vss within 2-fold error (FE).

Comparing the use of predicted versus experimental PL values in the modelling of Vss shows that, in 10 out of 12 models where PL was available during model training, the predictive accuracy was higher with pPL, i.e. experimental data complemented with the predicted phospholipidosis. Note that the compound set is the same for the dataset annotated with pPL or ePL (with the presence of about two-thirds of missing data being the

only difference between datasets), and the distribution of Vss values associated with missing PL does not vary considerably from that of measured PL (see Appendix III, Figure A3.1). Therefore, the increase in predictive performance from using pPL is due to an increase in information content brought by filling in missing PL values (otherwise ePL, which covers the same range of Vss values and is experimentally derived, is composed of higher-quality data, and would have produced higher quality decision splits). This also opens a new opportunity to enhance the modelling of Vss, since the completion of physiological variables with predicted responses is fully computational as it does not require any further experimental data, which means this could be applied in a high throughput context.

Table 6.4. Predictive performance of the two best models on the test set.

Model	conditions	Feature content	R^2	RMSE	MAE	GMFE	Within 2-FE	Within 3-FE
8a	RF-GA	FS-MDs + FS-PDs	**0.560**	**0.4497**	**0.3391**	**2.18**	56.0	**73.1**
16a	BT-GA	all features	0.529	0.4606	0.3453	2.21	**58.2**	71.6

6.3.2. Further Assessment of the Selected Model

The best model shows a good overall performance, with a coefficient of determination of 0.56 for the test set, which is close to expert recommendation ($R^2 > 0.6$) (Alexander et al., 2015). Furthermore, as seen by the scatter plot of predicted versus observed logVss (Figure 6.4.A) for the test set, the majority of external instances were predicted within a fold error of 3 (73.1%). Plotting the predicted against the observed logVss shows the characteristic tendency for underprediction near the upper limit of Vss, also observed in previous work (Freitas et al., 2015). By looking at the correlation shared between variables in the dataset, tested using a Spearman rank-order correlation test with Bonferroni correction, where statistically significant correlations, in red, are quite sparse. This does not mean that variables do not correlate, but rather, that they might hold local correlation amongst each other. This encourages tackling the problem at hand (predicting Vss) with machine learning, which allows harnessing such local relationships.

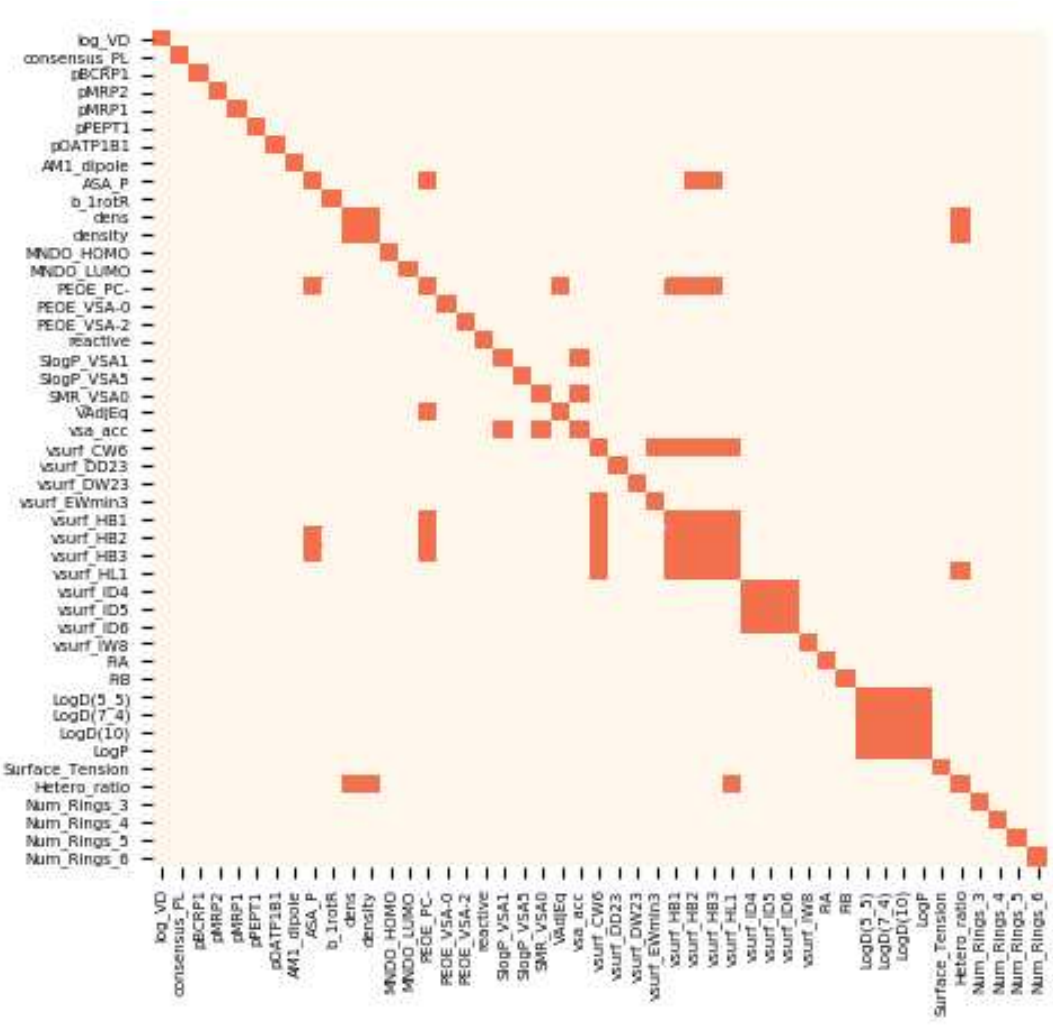


Figure 6.3. Correlation significance of all-against-all variables in the best model. Significant correlation between two variables is identified in red. This resulted from a Spearman rank-order correlation test with Bonferroni correction.

Looking at Figure 6.4.A it is possible to identify four evident outliers, which are labelled in the Figure's legend. Among these, pentamidine showed the highest error, which can be attributed to extensive and strong binding to lysosomes and extensive phospholipidosis (Filippone et al., 2011) with a resulting extensive tissue deposit (WHO, 2013). Although phospholipidosis is present in the model as a feature, this is clearly not sufficient for such extreme cases. One reason for this is that phospholipidosis data used here (and the only type of data available in sufficiently large amount for modelling) is binary yes/no in nature. This ignores the potency of drugs in causing phospholipidosis and hence the extent of the effect on Vss of individual drugs. Consequently, this limits the ability in capturing and predicting high Vd values due to extensive PL. Similarly, chloroquine also exhibits extensive tissue binding attributed to extensive phospholipidosis (lysosomal entrapment) (Zheng et al., 2011). Risedronic acid is known to be trapped in the bone (Watts and Diab, 2010),

resulting in high Vss. In the case of tigecycline the evidence is limited as this is a relatively more recent drug, a tetracycline analogue. It undergoes extensive tissue partitioning and has been reported to reach more than 20 times higher intracellular concentration than extracellular when a specific cell type (polymorphonuclear neutrophils) was studied (Ong et al., 2005). This cell accumulation as well as the formation of complexes with metal ions (Barbour et al., 2009) seems to contribute to its much higher volume of distribution than structurally similar tetracyclines.

Even though this model accounts for such effects as phospholipidosis or protein-mediated transport, these four compounds were still largely mispredicted. One main reason discussed earlier is that all experimental data in the physiological features is binary, which eliminates the ability to account for the extent of the physiological properties such as extent of phospholipidosis or the propensity of being transported by specific substrates. Hence, these parameters, although very useful as seen in the reported results, may prove inadequate in case of drugs with strong specific interactions with transporters or extreme cases of phospholipidosis. One other reason for such mispredictions can be a lack of sufficient chemical coverage in specific regions of the Vss response, which hinders the ability to learn local structure-response relationships. Examples of this are risedronic acid and tigecycline. Looking into the region of logVss that they occupy (i.e. [0.75, 1.15]) (Appendix III, Figure A3.2) shows that the training span of some of the main descriptors does not cover these two outliers. During training, this range of logVss was only covered by FiA values in the range [0, 0.134], and as a result compounds taking these FiA values were located closer to the unity line (i.e. above the outlier line in Figure 6.4.B), while these two outlier compounds have an FiA of 1, outside the training range. A similar situation occurs with vsurf_CW6. Considering that these features are both among the top 10 most important features, it is understandable why these compounds are mispredicted. Additionally, the other 5 of the top 10 features have values out of the training range in at least one of the outliers.

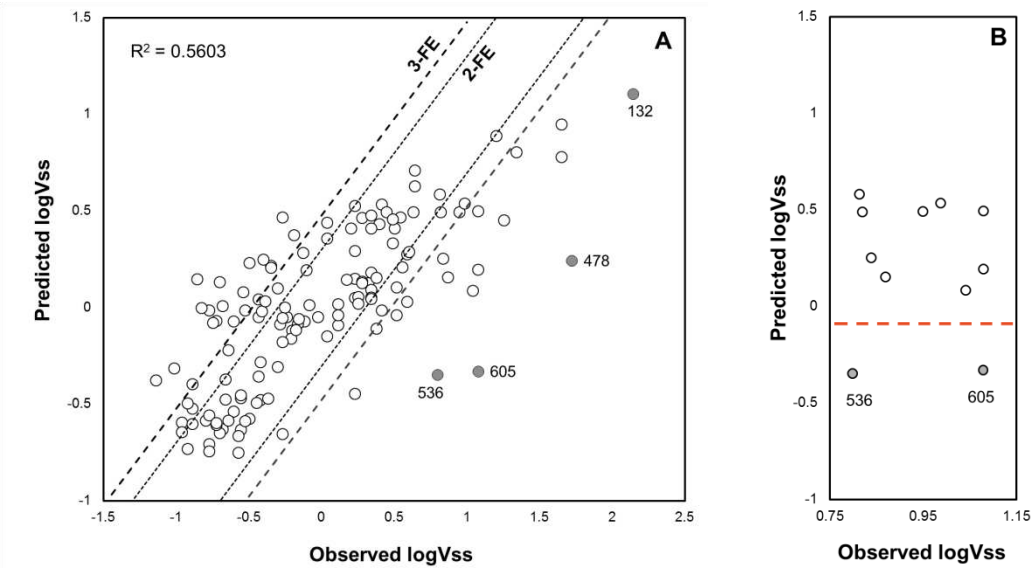


Figure 6.4. A) Predicted LogVss versus Observed logVss regression plot of the best model (8a). Four evident outliers are highlighted in grey (132: chloroquine; 478: Pentamidine; 536: Risendronic acid; 605: Tigecycline). B) Highlight of the region occupied by two outliers: 536 and 605.

The applicability domain profile built for this model (Figure 6.5) establishes a very robust relationship between predictive error and the level of disagreement (STD score) amongst the trees in the random forest of model 8a. In fact, all four outliers previously discussed would have been identified, in a real prospective testing scenario, as low confidence predictions, as they are only covered at STD values of at least 0.05, which is relatively large considering it coincides with the point at which the last third of instances start being covered.

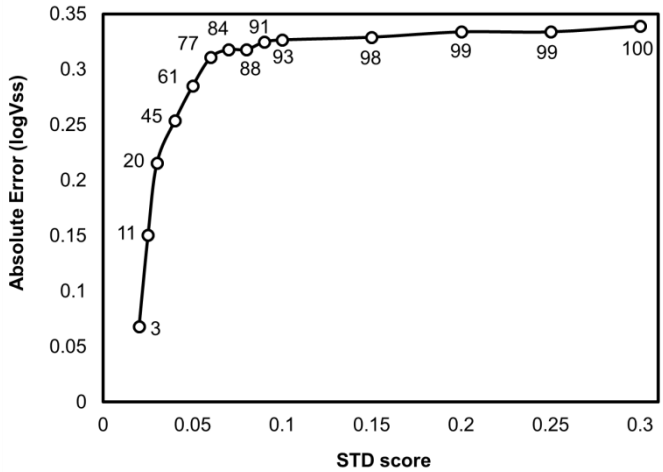


Figure 6.5. Applicability domain profile of model 8a. The data points are annotated with the percentage of the test data that is being covered as the AD limits are relaxed (i.e. the STD score increases).

Regarding the chemical content in the model, as expected, the top features in model 8a carry information on ionization state and lipophilicity (FiA, logD(10), log(7.4) and FiB), which have been widely implicated as determinants of the Vss (Ghafourian et al., 2006). In a correlation analysis between logVss and different molecular features, ionization state showed the strongest impact on logVss (Gleeson, 2008, Smith et al., 2015). As acids strongly bind to positively charged albumin, they are more prone to be confined to intravascular space; conversely, bases do not undergo such strong binding and instead have more affinity to membranes and tissue structures which contain negatively charged phospholipids (Smith et al., 2015, Gleeson, 2008, Zhivkova et al., 2015). Lipophilicity (clogP) combined with ionization (neutral and basic) state have also been correlated to distribution (Gleeson, 2008, Zhivkova et al., 2015, Ghafourian et al., 2006). However, given that the relationship between chemistry and the volume of distribution has already been extensively explored with several prior QSAR works on this same dataset and others (Zhivkova et al., 2015, Berellini et al., 2009, Ghafourian et al., 2006, Freitas et al., 2015, Lu et al., 2016, del Amo et al., 2013), this study will focus on the quantitative and qualitative impact of the different physiological features that have been selected into the final model, which is the main novelty aspect being explored here.

Recall that the best model (8a) was preceded by two parallel steps of feature selection performed separately on the PDs and MDs (and then the two sets of selected features were merged). This means that only six PDs were provided to the modelling step, however all six were selected into the model (see Appendix III, Table A3.6). From these, the physiological descriptor that shows the highest importance (as it participates in the modelling of the largest number of instances) is pPEPT1, implicated in the modelling of 24.5% of the instances. This is relatively high compared to the 47.1% affected by the most important descriptor (FiA) in the random forest that constitutes model 8a. Other physiological descriptors selected into the model were pPL, pMRP1, pMRP2, pBCRP1 and pOATP1B1. Looking into the content of the top nodes across the trees in the forest provides additional information on the importance of features in a given modelling task (Freitas, 2013), and in the random forest (1000 trees) constituting model 8a all the previously mentioned physiological descriptors occupy the top node in at least one tree, ranging up to 26 trees (see Appendix III, Table A3.9).

There are examples of the direct impact of each of these transporters in the volume of distribution of drugs (Dantzig et al., 2004, Grover and Benet, 2009). It is not clear why PEPT1 is the top physiological feature, but it is not unreasonable, however, to attribute this to the fact that PEPT1 is located in a variety of tissues; whereas, for example, OATP1B1 (the least important PD) is regarded as liver-exclusive (see tissue content summary in

Appendix II, Table A2.8). In addition to any possible biological explanation, the quality of the output (i.e. predictive accuracy) will likely influence the use of these features in the random forest. Note that all types of available sources of physiological information (i.e. phospholipidosis, ABC transport and SLC transport) are present in this model.

To ascertain whether these different features provide complementary information or are redundant as predictors of Vss, their presence in the random forest model was analysed in more detail. As seen in Table 6.5, a maximum of 5 (out of 6 available) transporters present in the same rule were observed across 56 rules containing 5 transporters as predictors. These 5-transporter rules were found in five different combinations (note that a given combination might occur in different rules, and might be accompanied by molecular descriptors). Table 6.5 shows that information of physiological nature (either isolated or in combinations) was used in 64% of the total set of rules in this random forest model, which also supports the value of accounting for transport information as well as drug-induced PL.

An analysis of the distribution of logVss values of the final leaf nodes affected by each combination of physiological descriptors was also carried out, and the results are shown in Appendix III, Figure A3.3 and Table A3.10. Overall a great diversity of rule combinations associated with different median Vss indicates a large degree of polyvalence, where different combinations of different features are able to cater to different locations in the Vd chemical space. Two examples of this are the two most extreme cases, combinations AU (logVss = - 0.665) and D (logVss = 0.460), where they share three features (MRP2, BCRP1 and OATP1B1), out of a total of 4 and 5, respectively.

Table 6.5. Frequency of combination sizes of physiological descriptors occurring in the same rule. The rules where these combinations occur may or may not contain molecular descriptors as well. 64% of the full collection of if-then rules contain at least one PD.

Number of PDs in combination	Rule count	% of full set of rules
1	88969	41.8
2	38250	18.0
3	8172	3.8
4	950	0.4
5	56	0.03
		Total = 64%

Finally, to identify if there is any correlation between the content of the combinations and the median Vss, the different combinations were plotted using multidimensional scaling to visualize the relative proximity between different combinations and their Vss simultaneously (Figure 6.6).

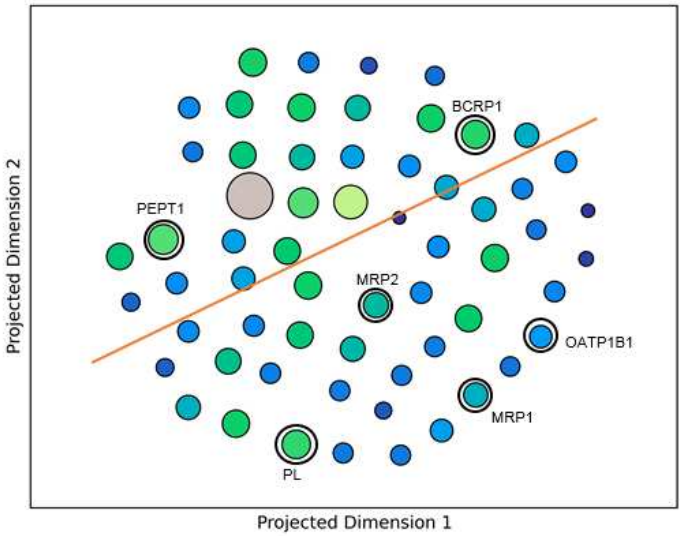


Figure 6.6. Visualization of proximity between the 61 unique label combinations (listed in Appendix III, Table A3.10) using t-SNE multidimensional scaling, where each combination is transformed into a binary vector where 1 represents the presence of a label and 0 its absence. This is done with respect to a total of 6 different features found in the full RF model (8a). Only the single-label combinations have been annotated in the figure, as a way to identify the relative locations of the features in the plot. Note that the plot does not represent absolute distances, but rather relative distances. Contrary to all other single-label combinations, MRP2 is not at the edges of the plot, which can be attributed to it being present in more combinations than all the others. Ascending values of logVss are portrayed from small (blue) to larger (green) circles.

To generate this plot, all unique combinations (N=61), listed in Appendix III, Table A3.10, were converted into a binary "on/off" vector where each bit corresponds to one of the 6 PDs. For example, the combination {pPL + pPEPT1} is represented as {0,1,0,0,0,1} where the 2nd and 6th locations are "on" as they correspond to the place of pPL and pPEPT1 for all combinations. This set of 6-bit vectors was submitted to t-SNE projection which attempts to arrange combinations according to their relative similarity (so absolute locations should not be interpreted). Larger (and greener/light brown) points indicate larger median Vss, whereas smaller and blue points indicate smaller median Vss.

There appears to be a separation between larger Vss, found in the vicinity of both PEPT1 and BCRP1, from smaller Vss. Such separation can be tentatively demonstrated using a visually-derived separation line (orange). This could indicate that combinations of physiological features containing PEPT1 + BCRP1 (and MRP2 to some extent) tend to be present in prediction rules handling higher Vss, as opposed to rules containing combinations of PL, MRP1 or OATP1B1.

To further validate the presence of the physiological descriptors in the best model, the feature set provided to train model 8a (which results from feature selection using GA search) were used to train a simpler M5 model tree. This has the main purpose of challenging the

possibility of the various PDs being present in the model only by chance, as an effect of training a very large ensemble (1000 trees) of unpruned trees. As shown in Table 6.6, the M5 model tree using CFS-GA-selected features produced 4 rules containing 3 out of the 6 PDs provided to modelling. Additionally, when providing the entire feature set (MDs and PDs) to training, the obtained M5 model also contained PDs (3 out of 11) (Table 6.6). Both observations show that PDs are deemed as valuable predictors even in a much simpler model with a single model tree. Note that, regarding the M5 model obtained using all features, the 11 PDs competed against 304 MDs, and in a model comprised of one single rule 3 PDs were selected among 47 total descriptors. If the information carried by these selected PDs was redundant with respect to MDs, the chance of the former to be picked into the model would be considerably smaller than the latter.

Table 6.6. Summary of the results obtained by the M5 model tree built with different sets of descriptors. The best performance values are highlighted in bold and underlined. This exercise tests the ability of PDs being selected in a harsher embedded feature selection environment, and is not meant to create alternative (competitive) models to the RF and BT models.

	All descriptors (11 PDs + 304 MDs)	descriptors selected by CFS-GA (6 PDs + 41 MDs)
R^2	0.445	**0.469**
MAE	0.368	**0.338**
RMSE	0.478	**0.461**
	1 rule; pBCRP1, pOATP2B1, pPEPT1	4 rules; pMRP2, pPEPT1, pPL

6.3.3. Comparison with Other Works on Vd Modelling

The dataset used in this work is regularly used as a benchmark for the modelling of volume of distribution (del Amo et al., 2013, Gombar and Hall, 2013, Zhivkova et al., 2015, Zhivkova and Doytchinova, 2012, Demir-Kavuk et al., 2011), which allows a fairer comparison between different works.

Del Amo et al (del Amo et al., 2013) opted to remove challenging compounds from this dataset (i.e. 4 biphosphonates which are known to accumulate in the bone tissue, and 2 anti-malarials), whereas in this study these compounds were kept to challenge the modelling exercise and keep the modelling conditions as close as possible to a real world scenario. Additionally, in their work external testing is either performed on compounds evenly sampled within the model's chemical space, or compounds that are found inside the applicability domain boundaries. It is likely that this might have produced an overoptimistic predictive accuracy.

Gombar and Hall (Gombar and Hall, 2013) used this same dataset but removed a considerable number of compounds that were found outside a chemical space limit, and have used test compounds that were not sampled from Obach dataset. Some of these external test compounds came from Berellini et al (Berellini et al., 2009), who modelled Obach dataset as well. They allocated the full dataset used in this work into the training set, and tested the predictive performance on a very limited dataset of 29 compounds, collected outside the Obach dataset.

Zhikova and colleagues used the Obach dataset, but they only modelled basic drugs (Zhivkova et al., 2015) or acidic drugs (Zhivkova and Doytchinova, 2012). The argument of modelling acids and bases separately in order to properly address their different distribution patterns (Zhivkova et al., 2015, Zhivkova and Doytchinova, 2012) is not applicable if regression algorithms that can naturally create such partitions are used, like the tree-based methods used here. In fact, this is precisely what was observed in the results, where the best model showed that many trees within the random forest have, as top nodes, features that explicitly characterize ionization type (FiA and FiB) (see Appendix III, Table A3.9). Additionally, past studies have demonstrated that this approach of modelling acids and bases separately does not yield improved predictive performance (Ghafourian et al., 2004, Ghafourian et al., 2006), So, the full available data can be modelled as a whole, which also has the advantage of allowing one to account for common features between different chemical groups such as acids and bases.

Demir-Kavuk et al.(Demir-Kavuk et al., 2011) removed 86 compounds from this dataset, for which any of the descriptors could not be calculated. This will likely limit the applicability of the model. Still, despite the removal of so many compounds, their best model obtained a GMFE of 2.08 in an external test set which is slightly superior to the best model (2.18) in this work, probably due to their decision towards a more optimistic modelling of Vss (i.e. modelling under less challenging conditions, with the removal of compounds from training and test sets that are more difficult to parameterize in descriptor calculation software).

The study reported by Freitas et al. (Freitas et al., 2015) has used the most similar approach to the current work, to date. In this work, predicted tissue partition coefficients were introduced as descriptors into the modelling of Vss. Similarly to the findings observed here, their work showed a small improvement in the ability to model Vss when some physiological features were introduced into the modelling. Although the source of physiological input (tissue:plasma partition coefficients) was very different from the sources used in this work, the overall modelling process was relatively similar to this one in terms of random allocation of compounds to testing, the same sources of molecular descriptors, and similar type of

tree-based ensemble algorithm. While their best model had a GMFE of 2.29, the current best model had a GMFE of 2.18, which is slightly better and might indicate an improvement in predictive accuracy when providing transporter and phospholipidosis information to the modelling process.

Louis and Agrawal (Louis and Agrawal, 2014) used a much smaller dataset (97 training instances and 24 test instances) which is, with some exceptions, a part of the Obach dataset composed of a more limited Vss range ([-1, 1.32]) and annotated with descriptors of chemical nature only.

6.3.4. Benchmark Comparison on a Benchmark Test Set

There are two recent works (Gombar and Hall, 2013, Lombardo and Jing, 2016) that have tested their models in external test sets, and provided the complete set of predictions obtained, which allows direct comparison of the models' predictive power with this work. Gombar and Hall (Gombar and Hall, 2013) used the same dataset used here to build the QSAR models, whereas Lombardo and Jing (Lombardo and Jing, 2016) used a larger Vss dataset (N = 1096). Comparing this work's predictive performance against that of Gombar and Hall (scenario 1) allows assessing the value of introducing physiological information into modelling, as this is the major difference between both modelling routines – other secondary changes are present, like the removal of problematic compounds, however these are considered minor. On the other hand, a comparison against Lombardo and Jing (Lombardo and Jing, 2016) (scenario 2) allows determining the value of increasing chemical space through the increase in observation count (quantitative improvement) versus providing enriched input through the addition of physiological information (qualitative improvement).

In scenario 1, the best model in this chapter (8a) showed improved performance in all calculated measures for an external set of 30 compounds taken from Gombar et al (Gombar and Hall, 2013), as shown in Table 6.7 and Appendix III, Figure A3.4. Considering the smaller training set size used in this study (Gombar and Hall: N = 569; this work: N = 398), the better performance of model 8a in external prediction is notable (other things being equal, more data should lead to better performance). As a result, the superior performance of model 8a may be attributed to the availability of physiological input during training, or the modelling scheme used in this work which are the major differences between both works.

Table 6.7. Summary of predictive performance from Gombar and Hall (Gombar and Hall, 2013) and this work (model 8a), evaluated on an external dataset (N = 30).

		models in (Gombar and Hall, 2013)	
	m_8a (this work)	SVM	MLR
MAE	**0.205**	0.264	0.422
GMFE	**1.604**	1.835	2.641
MFE	**1.869**	1.995	5.430

Regarding scenario 2, with the external set of 34 compounds obtained from Lombardo and Jing (Lombardo and Jing, 2016), it can be observed that despite the fact that model 8a was trained with a significantly smaller training set (N=398 versus N=1096), it is still able to show comparable performance to other models as seen in Table 6.8. This model could also overcome some extreme mispredictions which were still mispredicted by model 8a, but to a lesser extent (See Appendix III, Figure A3.5. to compare the plotted observed vs predicted for models in Table 6.8). Furthermore, 53% of model 8a's predictions show smaller error than RF_33 (which is the model with the smaller MAE value in Lombardo and Jing (Lombardo and Jing, 2016)). This supports the validity of the selected (best) model and may also indicate the value brought by accounting for physiological processes.

To further determine the value of using physiological features in the modelling of Vss in a larger chemical space, the best model from both this chapter and that from Lombardo et al. were trained with Lombardo's entire dataset of 1096 compounds and the resulting models were used for the prediction of the external test set of scenario 2. Note that the modelling algorithm used in both models is random forest. The difference between Lombardo's model and model 8a is the features used in the analysis (different Volsurf+ molecular descriptors). In addition, different algorithm parameters were used as per the original study, i.e. no minimum node size set for this work's model and a minimum node size of 10 for Lombardo et al, and descriptor sampling per split set to WEKA's default for the current model versus 11 for Lombardo et al. To test the effect of physiological descriptors, both models were tested with and without the presence of these descriptors. Still, despite the different conditions, using both the current set of parameters and Lombardo et al's set of parameters yielded the same conclusion: including PDs improves predictive performance across all measures, as summarized in Table 6.9.

Table 6.8. Summary of predictive performance measures from Lombardo and Jing (Lombardo and Jing, 2016) and this work, evaluated on an external dataset (N = 34).

	models in (Lombardo and Jing, 2016)

	m_8a (this work)	RF_33	PLS_11	consensus RF_33 and PLS_11
MAE	0.305	**0.302**	0.363	0.317
GMFE	2.017	**2.003**	2.308	2.073
MFE	**2.276**	2.300	2.970	2.510

Lastly, it should be noted that the difference between retrained m_8a and the retrained Lombardo's model can be explained by the fact that the latter was selected in the original publication as the best (standalone) model based on the performance obtained on this same test set. As a result, comparing both models is not fair as Lombardo's model is bound to be superior for this particular test set, hence why this study focuses on comparing presence or absence of PDs within each model.

Table 6.9. Summary of predictive performances from the different variants of the Vss modelling conditions. All performances result from testing the models on a fixed, common dataset.

		Retrained m_8a model		Retrained Lombardo's model	
	m_8a (this work)	MDs only	MDs & PDs	MDs only	MDs & PDs
MAE	0.305	0.322	**0.318**	0.300	**0.293**
GMFE	2.017	2.104	**2.080**	1.993	**1.962**
MFE	2.276	2.728	**2.689**	2.290	**2.253**
Number of predictions with the smallest error	**8**	7	7	4	5

Lastly, it should be noted that, surprisingly, model 8a was the one generating the highest rate of the smallest prediction errors (out of all 5 models in Table 6.9), which means that it shows the highest number of predictions associated with the smallest error across all 5 alternative models.

There is an alternative theoretical hypothesis that transport holds no significant additional value based on the fact that many correct Vss predictions are made from compounds that undergo protein-mediated transport (Berellini et al., 2009). However, the impact of transport may vary across compounds, and a given compound that is transported and generates a 2-fold error is perceived as being correctly predicted. Perhaps accounting for the transport effect in this case would reduce the error from 2-fold to closer to 1 (perfect prediction). Indeed, this is what the current work demonstrates, whereby the addition of physiological information to the modelling improves the model's performance in a systematic manner.

6.4. Conclusions

Modeling distribution using only the chemical information of compounds has proven difficult, since such an approach does not successfully account for the specific interactions between drugs and the physiological system that govern Vd. When modelling Vss, mispredictions are generally attributed to transport or tissue binding processes, and this is to some extent the general assumption even for unexplainable mispredictions, as seen in the literature (del Amo et al., 2013, Lombardo and Jing, 2016). This demonstrates the importance of addressing transporters in modelling drug distribution.

This chapter explored the impact of using key physiological processes as input information in the modelling of human Vss. However, as descriptors of this nature are obtained experimentally, it was proposed that physiological features could be modelled in a prior step (some of which was done in chapters 4 and 5), and the learned (predicted) responses would be used to complete the data on experimental responses. At the limit, this could potentially be used as the standalone source of physiological information. The physiological parameters used in this work capture information about the potential of drugs to be transported by ABC or SLC transporters (substrate/non-substrate data) and the potential of drugs to accumulate in tissues through drug induced phospholipidosis (again a categorical variable).

It was observed that, across different variations of regression methods or feature selection techniques, adding physiological descriptors improves the predictive performance of Vss in the great majority of cases. Additionally, it was observed that using predicted physiological data to fill in missing experimental observations, specifically regarding phospholipidosis, improved the predictive performance in the majority of cases, when compared to using just experimentally observed PL responses.

To validate the main premise of this chapter that physiological descriptors are useful features in Vd modelling, the best model obtained in this study was compared to: (1) a model built on the same dataset as the one used here, and (2) a model built on a considerably larger dataset, both only using molecular descriptors. Direct comparisons were possible through testing on two relatively small external datasets (one used by each of the mentioned models), which revealed that the best model in this work performed better than, or similarly to, previous models, and the incorporation of physiological descriptors improves models obtained by both methods.

The work presented in this chapter not only shows the value of using transporter and phospholipidosis data as input descriptors for the modelling of the Vd, but also opens a

precedent for the possibility of predicting physiological responses and using those predictions to complete missing data, in order to aid the learning of the Vd QSAR model.

7. Accounting for Transporter Binding, Transporter Tissue Expression, Phospholipidosis and Plasma Protein Binding in the Modelling of Volume of Distribution

7.1. Introduction

The main premise of Chapter 6 was that, in order to better address the modelling of volume of Distribution (Vd), there is a need to account for the interplay between the specific and unspecific components of distribution, which are driven by both physiological and chemical factors. Chapter 6 approached this problem by considering the net effects of these factors across the entire human body, handled as a whole. However, this has the limitation of not considering that such determinants vary across the full collection of tissues in the body. In light of the recent availability of expression profiles of a wide range of proteins across a range of tissues, available with the Human Protein Atlas (Uhlén et al., 2015), it has been hypothesized that the information on the expression profiles of key transporters that drive distribution can be used to help predicting Vd.

As the transporters considered in Chapter 6 are expressed to different extents in different tissues, as discussed in the Introduction Chapter 1, and reiterated with the data from the Human Protein Atlas, taking into account tissue expression could, theoretically, refine the information carried by the ABC and SLC binding descriptors used in chapter 6. Additionally, another extension of the approach used in the previous 6 is the incorporation of plasma protein binding as one of the descriptors. This feature has been considered to be among the physiological factors with the largest, direct influence over Vd (Curry and Whelpton, 2017).

Such wealth of features of physiological nature combined for the modelling of Vd is unprecedented, as previous works only went as far as using using tissue partition coefficients in Vd modelling (Freitas et al., 2015, Paixão et al., 2014), plasma protein binding or membrane binding (Sui et al., 2009, Hollósy et al., 2006).

7.2. Methods

7.2.1. Volume of Distribution Dataset

The dataset described in Section 6.2.1 was also used in this chapter. However, contrarily to Chapter 6, here plasma protein binding (PPB) was also used as an additional physiological descriptor. The final dataset used for the modelling consisted of log-transformed Vss (logVss), 304 molecular descriptors (MDs) and 11 physiological descriptors (PDs): PPB, drug-induced phospholipidosis (PL), 4 PDs referring to ABC efflux (mediated by P-gp, MRP2, BCRP and MRP1) and 5 PDs referring to SLC uptake (mediated by OCT1, PEPT1, OATP1B1, OATP1B3, and OATP2B1). In chapter 6 OATP1A2 was used as a physiological descriptor. However, as the purpose of this work is to explore the effect of accounting for transporter expression levels, and OATP1A2 did not have any reported expression data in the used source (see Section 7.2.2) at the time this study was carried out, this was excluded as a descriptor. Recall that, as explained in Section 3.1.3 in the general Methods Chapter as well the methods section 6.2.1 of the previous chapter, missing data from the PDs is filled in with the respective predictions. As done in chapter 6, these PDs where predicted responses were added to the pre-existent experimental data are prefixed with a “p”.

7.2.2. Tissue Expression for the Correction of Transporter Data

In order to refine the transporter descriptors, transport data was corrected with expression levels of the ABC and SLC transporters, which were gathered from The Human Protein Atlas platform (www.proteinatlas.org). These are derived from high quality direct quantification through western blot. Expression levels are reported in four main qualitative levels: high, medium, low and not detected. These were converted into fraction equivalents, namely, 1, 0.6, 0.3, and 0, respectively. Three different expression correction schemes were tried, as explained below. The corrected PDs are suffixed with “_c”, standing for “corrected” (done in addition to the “p” prefix).

Scheme #1: Each transporter response (R) was corrected by being multiplied by the sum, across all available tissues, of the product of the corresponding expression level (E_t) and the tissue weight (W_t), as shown in Equation 7.1, where t indexes a tissue.

$$R \times \sum_t E_t \times W_t \qquad \text{(Eq. 7.1)}$$

Scheme #2: To explicitly test the idea of transporters contributing as a whole to produce the observed Vss, all transporter values obtained from scheme #1 were summed into a single feature (a more abstract descriptor), for each compound in the dataset. This new feature (named “distribution”) can be regarded as the global transport effect, and was used in place of the original transporter features.

Scheme #3: To further refine the impact of different transporters towards Vss, efflux towards excretion is differentiated from efflux towards interstitial space and/or systemic circulation. The first efflux scenario will be attributed a negative penalty p = -1, while the second will be assigned a negative penalty p = -0.286, which corresponds to the ratio of interstitial volume to intracellular volume (12 : 42 L) (Equation 7.3).

$$\begin{cases} R \times \sum_t E_t \times W_t & \text{for uptake} \\ R \times \sum_t -p \times E_t \times W_t & \text{for efflux} \end{cases} \qquad \text{(Eq. 7.3)}$$

To address the possibility of the tissue weights spanning across 3 orders or magnitude, all schemes described above were repeated with scaling of tissue weights, by applying a log transformation. In preliminar analyses, given the two best model produced (based on the validation set MAE) were achieved with no tissue weight scaling, the use of no scaling was selected as the optimal setting for modelling and became the focus of this chapter’s discussion.

7.2.3. QSAR model building

In chapter 6 boosted regression trees (BRT) and random forests (RF) were the two regression algorithms tested, both paired with genetic algorithm search (GA) and greedy stepwise search (GS) feature selection (FS) algorithms. Since the dataset remains the same, and the best overall combination was RF paired with GA, these conditions were maintained here. As a result, to build the QSAR models, the GA pre-processing step was carried out by re-running the search 10 times using variable random seeds, and picking the features that were selected in at least 5 of those 10 runs. All feature selection parameters are described in section 3.3. The final feature set was then fed into the RF regressor, which was tuned using 10-fold cross validation applied on the training set, where the number of trees was optimized in a range between 100 and 1000 (at increments of 100). Different variations of feature sets were tested, namely: all features submitted to FS (FS-All), PDs

used directly (PDs) or submitted to FS (FS-PDs), and two separate feature selection routines applied to MDs and PDs separately, and merged afterwards (FS-MDs + FS-PDs).

This means that, for conditions containing "PDs", scheme #1 and #3 models had access to 10 PDs (9 transporter variables + 1 PL) while scheme #2 models had access to 2 PDs ("distribution" + PL).

All pre-processing and model training was carried using WEKA version 3.8 (Hall et al., 2009).

7.2.4. Retraining the Best Model with Plasma Protein Binding and Using a Larger Dataset

Upon selection of the best model, this was retrained with an additional physiological descriptor – PPB. This feature has been widely accepted as a predictor of Vss, and it was introduced as a feature to test whether providing it to the modelling step using the same procedure used for the remaining physiological features in this work would bring any value to the modelling of Vss. To make the problem more challenging (so as to avoid an overly optimistic scenario), a situation of sparse data availability was simulated by ignoring the available PPB data provided in the Obach dataset, and using the AstraZeneca dataset instead, which was also used to build a QSAR model from which prediction for missing PPB data were obtained(see Section 3.1.3). This also allows for a larger dataset from which to train a predictive PPB model, thus maximizing the chances for a better performing model.

Additionally, the best model conditions were reapplied to a larger, more recent Vss dataset published by Lombardo and Jing (referred to as the "Lombardo dataset" from this point onwards). To do this, a pre-processing feature selection step using GA was applied to this dataset, following the previously described procedure, and the resulting feature set was used to train a random forest model. The modelling conditions were optimized using 10-fold cross validation using the same procedure applied for the other models in this work. As Lombardo and Jing also provided their modelling conditions, this allowed recreating their models for further assessment of the impact of using physiological descriptors. As this same procedure was applied in the previous chapter, for more details on the processing of this dataset and the applied modelling conditions, the reader can refer to Section 6.2.3.

7.2.5. Comparison against Previous Models Using a Benchmark External Set

In order to challenge the value of adding physiological descriptors, this work's best model was compared to two other works by Lombardo and Jing (Lombardo and Jing, 2016) and

Gombar and Hall (Gombar and Hall, 2013). As described in the previous section, the former was obtained from a larger Vd dataset, while the latter was obtained from the same dataset as used in this work (Obach dataset). This comparison was done through two external datasets, used by each publication respectively.

7.2.1. Model Evaluation and Validation

Measures of predictive performance such as the coefficient of determination (R^2), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Geometric Mean Fold Error (GMFE), calculated as shown in Section 3.5.2, were used to compare the different modelling conditions tested. All comparisons were done based on the validation set. To select the best candidate model, all different models were compared using the validation set performance, and the two best modelling schemes were selected based on the lowest MAE. In addition, the percentage of predictions within 2- and 3-fold error (FE) were used to compare the two best candidate models. Upon selecting these, they are compared based on the test set and the best model is finally selected for further evaluation.

To validate the best model, its applicability domain was also characterized by using the standard deviation of an ensemble of predictions (named STD) as a reliability measure (described in Section 3.6). t-SNE muldimensional scaling was used for visualization.

7.3. Results and Discussion

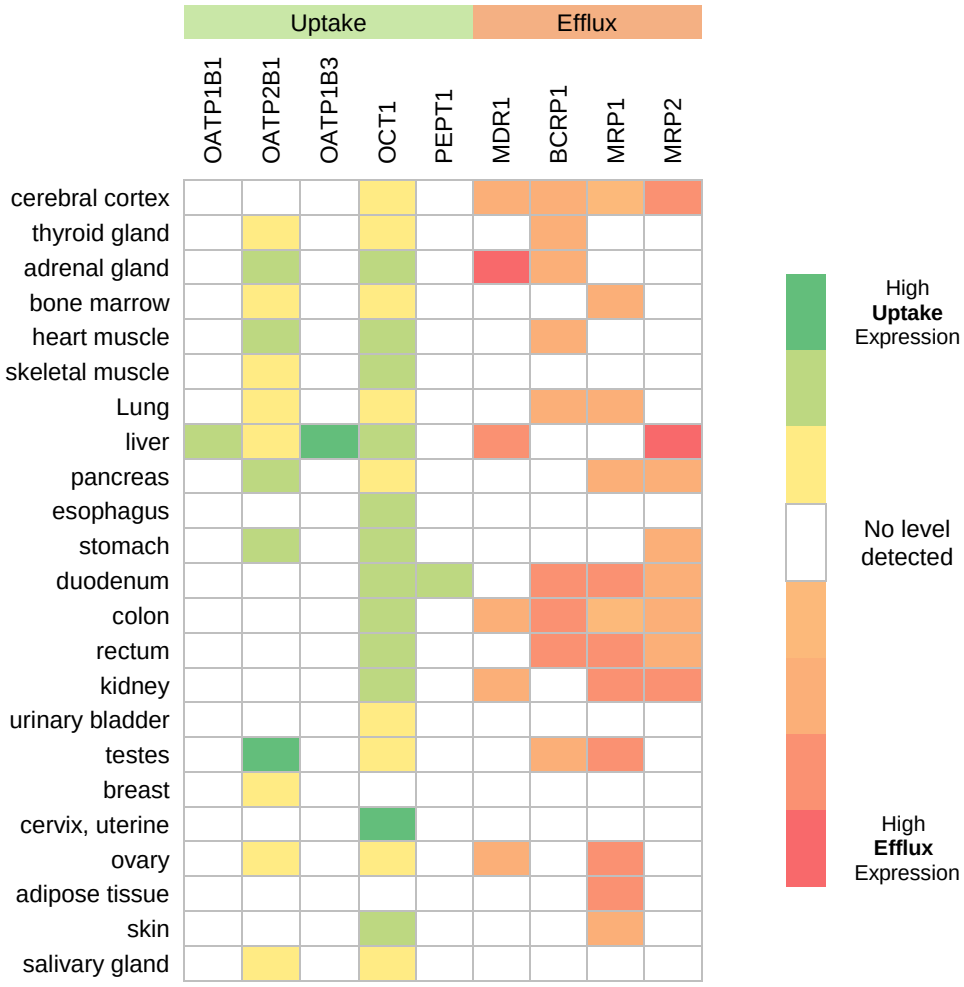
7.3.1. Overall Evaluation of Model Performance

Following the work in Chapter 6 where protein-mediated transport and drug-induced PL were found to be useful predictors of Vss, and the interaction of compounds with several transporters had an important impact on the final model, this current chapter aims at exploring further the impact of transporter interactions by applying corrections based on tissue expression levels of these transporters. To do so, the measured protein expression levels across a range of tissues were used to refine the information content provided by the likelihood for efflux and uptake of compounds (expressed as a probability that spans between 0, for extremely likely non-substrates, and 1, for extremely likely substrates).

The landscape of relative expression levels of transporters across tissues is represented in Figure 7.1. This information was used to correct the transport data, where expression was used in its absolute values, as well as relative values (adjusted for direction of transport), as detailed in Section 7.2. From the four expression correction schemes tested, scheme #3

seemed the most promising one as it is the most physiologically accurate, being the only which accounts for the location of transporters in the cell membrane (i.e. apical or basolateral side) across different tissues. Surprisingly, scheme #3 did not produce any of the 2 best models (Table 7.1).

Figure 7.1. Expression levels of the transporters used in this work across a range of tissues, retrieved from the Human Proteine Atlas project.



Furthermore, for models derived from feature selection applied to the entirety of available descriptors (M1, M5, M9 and M12), all transporter expression correction schemes except scheme #3 marginally reduced the MAE relative to the best baseline model (i.e. best no-correction model, see legend of Table 7.1 for details). This might be an indication that distinguishing between efflux towards excretory fluids and efflux towards the blood is not a useful correction approach for transporter data (at least not in the way it was done here).

Table 7.1. Summary of the internal validation performance of the various modelling conditions tested. They are compared with the best model obtained in Chapter 6 (named there as model 8a), here identified as "best previous" in this current Table. All models under the same modelling block as the best model (i.e. regression + feature selection conditions that produced the best model) are here considered as the baseline, and identified as such in this Table. Here, "baseline" means the best scenario using no form of transporter expression correction. The two best models, selected based on the lowest MAE are highlighted in boldface. The downward arrows indicate an improvement against the equivalent baseline (no correction) models.

Expression Correction Scheme		no expression correction (BASELINE)						
				MAE	R^2	RMSE	**GMFE**	**Features Used**
None	M1	FS-All		0.3079	0.469	0.445	2.03	MDs, PDs
	M2	FS-PDs		0.4240	0.160	0.582	2.66	
	M3	PDs		0.3706	0.267	0.523	2.35	
	M4	**FS-MDs + FS-PDs (Best in Chapter 6)**		**0.3056**	0.474	0.442	2.02	MDs, PDs
		expression correction						
				MAE	R2	RMSE	**GMFE**	**PDs Used**
Scheme #1	**M5**	FS-All	↓	**0.3003**	0.5013	0.432	1.9967	3 transp.
	M6	FS-PDs	↑	0.4380	0.1352	0.5919	2.7416	
	M7	PDs	↑	0.3800	0.2586	0.5282	2.3988	
	M8	FS-PDs + FS-MDs	↑	0.3057	0.4758	0.4409	2.0216	4 transp. + pPL_c
Scheme #2	**M9**	FS-All	↓	**0.2954**	0.5057	0.4291	1.9742	pPL_c
	M10	PDs	↑	0.4716	0.0747	0.6225	2.9621	
	M11	PDs + FS-MDs	↓	0.3039	0.4818	0.4389	2.0133	distribution + pPL_c
Scheme #3	M12	FS-All	↑	0.3096	0.4732	0.4426	2.0399	4 transp.
	M13	FS-PDs	↑	0.4387	0.1354	0.5925	2.7460	
	M14	PDs	↑	0.3770	0.2637	0.5264	2.3823	
	M15	FS-PDs + FS-MDs	↑	0.3083	0.4698	0.4437	2.0338	4 transp. + pPL_c

The two best models produced here (M5 and M9) were both generated from previous feature selection procedures applied to all features simultaneously. Despite the overwhelming number of molecular descriptors competing against few physiological descriptors (10 and 2 physiological descriptors initially available for the modelling process – preprocessing and training – of M5 and M9, respectively), the latter were still selected into the final descriptor set, used for model building. In model M5 the physiological descriptors in the final set consisted of 3 expression-corrected transporter descriptors (pBCRP1_c, pMRP2_c and pOATP1B1_c with substrate/non-substrate data, see Table 7.3) and for model M9 only pPL was selected.

These two selected candidate models were tested on the (left-out) test set (0.3237 and 0.3286 for M5 and M9, respectively). Based on this, model M5 was further evaluated, as will be discussed below.

Regarding the validity of predictions in the test set, upon which one relies to draw conclusions about the value of using expression-corrected transport data to improve the ability to model Vss, Figure 7.2 shows that there is a robust correlation between the STD scores (which are here used as the predictive reliability measure) and predictive error in unseen data. This means that the model can be used prospectively with confidence, as the STD value serves as a good surrogate of relative expected error for predictions.

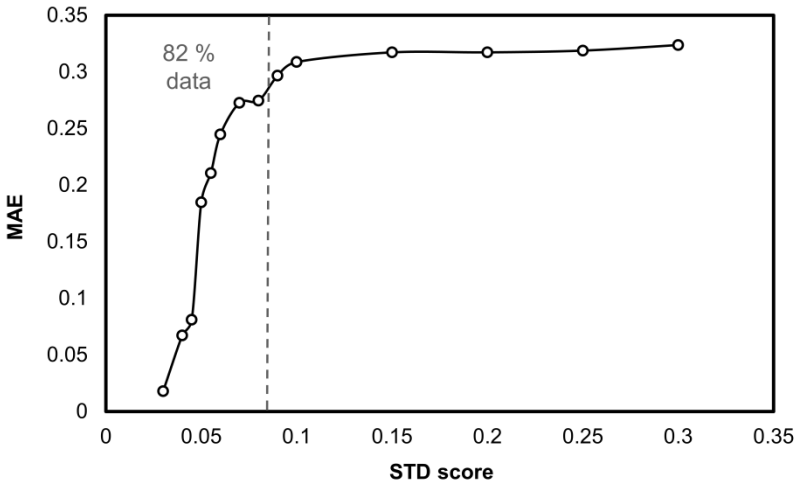


Figure 7.2. Applicability domain profile of model M5 (the best model on the test set). The test data is sorted according to their STD score, and their respective MAE values within increasing STD score threshold are recorded.

7.3.2. Impact of Expression-corrected Transporter Features

The best model with transporter-expression correction was generated from applying feature selection to the full set of features (M5), contrarily to the best model with no expression correction, which was trained with feature selection applied to both physiological descriptors and molecular descriptors, separately (M4). Despite being subject to feature selection alongside a much larger number of molecular features (304 molecular descriptors and 11 physioloical descriptors), some of the physiological descriptors were selected in the final subset (pBCRP1_c, pOATP1B1_c and pMRP2_c). Even though a relatively small percentage of compounds was affected by any of these three transporters, with pBCRP1_c showing the largest descriptor importance (13.9%), looking into the composition of the random forest model actually shows that both pMRP2_c and pOATP1B1_c were found at the top node in 16 and 4 trees, respectively (out of a total of 600 trees). Finding a descriptor close to (or at) the top node is a strong indicative of the meaningful role of such descriptor towards the modelled output variable (Freitas, 2013).

Table 7.2. Summary of predictive performance measured on the test set (N=134) for the best model in this work (M5) and the best model from Chapter 6 (8a).

Best Models	conditions	Feature content	R^2	RMSE	MAE	GMFE	MFE	Within 2-FE (%)	Within 3-FE (%)
8a	RF-GA	FS-MDs + FS-PDs	0.560	0.4497	0.3391	2.18	2.99	56.0	**73.1**
M5	RF-GA	FS-All features	**0.582**	**0.4353**	**0.3237**	**2.11**	**2.84**	**59.7**	72.4

Table 7.2 shows that using protein expression-corrected transporter data led to an improvement in the ability to predict Vss, showing an MAE of 0.3237 vs 0.3391 when no expression correction was applied (model 8a). Beyond MAE, all performance measures, except for % predictions within 3-FE, were better for the current chapter's best model, M5. Additionally, 58.2% of the prediction errors were reduced when compared to the best model trained in the absence of expression correction. As seen in Figure 7.3, these 58.2% (shown as the filled circles) include the instances associated with the largest prediction errors. In fact, almost all compounds in the right-hand side of the plot, which fall outside the 3-FE threshold (indicated by the dashed lines), are predicted with higher accuracy when using transporter-expression correction (when compared to model 8a). The improved predictions include four problematic compounds pointed out in previous works as well as in Section 6.3 as being particularly challenging to predict due to extensive binding to different tissue structures(WHO, 2013, Watts and Diab, 2010, Barbour et al., 2009, Zheng et al., 2011) (pentamidine, chloroquine, risedronic acid and tigecycline).

As the relationship between chemistry and Vss has already been extensively discussed in the literature, when discussing the chemical descriptors of the best model (M5) the focus will be placed on discussing the presence of physiological descriptors in the model. However, it should be pointed out that, besides the high importance in model M5 of expected descriptors such as ionized fraction and lipophilicity (which occupied the top positions in the random forest's trees in Chapter 6), a new descriptor implemented in MOE (h_pavgQ) has been found to have the highest feature importance in model M5. This is a descriptor calculated using Extended Hückel Theory, which is a semi-empirical quantum mechanics method that takes into account local resonance and electron withdrawing effects (Labute et al., 2014). In particular, h_pavgQ is the average total (formal) charge, a pH dependent parameter calculated based on the relative concentration of various protonation states of the molecule. This parameter conveys similar information to fractions of anionic, cationic, zwitterionic and unionized forms of a molecule at different pH values, which are calculated from the acidic and basic pKa values (Ghafourian et al., 2006). These fractions

were found to be major predictors of VD as distribution of compounds may be limited for acidic compounds such as nonsteroidal anti-inflammatory drugs with strong plasma protein binding, whereas basic compounds may be able to accumulate in the phospholipid membranes (Ghafourian et al., 2006, Freitas et al., 2015).

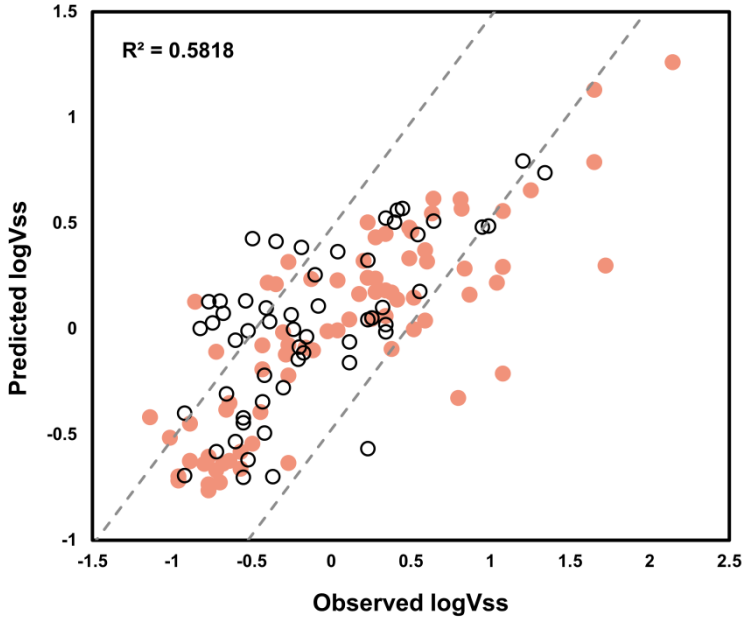


Figure 7.3. Scatter plot of test set predictions obtained by the best model with transporter-correction expression (model M5). Filled circles indicate predictions which show a smaller error compared to the best model with no transporter-expression correction in Chapter 6 (model 8a). Compare to Figure 6.4 to see the improvement for the outlier predictions.

In the previous attempt to model this same dataset without transport-expression correction in Chapter 6, even though exactly the same training set was used as well as the same feature selection and regression algorithms, there was a marked difference in the physiological descriptors that were selected into the model. In Chapter 6 and pPEPT1, pPL, pMRP1, pMRP2, pBCRP1 and pOATP1B1 were selected and used in the best achieved model (see Appendix IV Table A4.1 for full list of descriptors of model 8a), while in the current work the first three of these descriptors were not selected in the best model (M5). This difference is especially significant, as one of them (pPL) encodes information on experimental and predicted phospholipidosis (completed missing experimental data), and its absence associated with improved performance goes against the observations in Chapter 6, where pPL consistently improved the majority of the models produced. However, while the selection of a descriptor is evidence of its informative value, the failure to select a descriptor does not imply a lack of informative value. It might be the case that pPL was not selected due to the selection of other descriptors which were correlated with pPL or even

more informative than this descriptor (making the selection of pPL unnecessary), as M5 (contrarily to 8a) resulted from feature selection applied to the full set of features (making pPL redundant).

Table 7.3. Full list of descriptors used in the best model (M5), and their relative importance (in parenthesis), calculated as the percentage of correctly predicted training compounds, over the total number of training compounds that go through a decision node containing each of the descriptors in the model.

Descriptors in model M5	
h_pavgQ (42.2)	LogD(5_5) (16.5)
FiA (35.2)	chi1 (16.4)
LogD(10) (35.1)	vsurf_W7 (16)
vsurf_HL1 (26.4)	dipole (15.7)
FiB (25.3)	vsa_acc (15.1)
ASA_P (24.9)	vsurf_Wp5 (14.3)
vsa_pol (23.6)	vsurf_IW5 (14)
vsurf_HB2 (23)	**pBCRP1_c (13.9)**
LogP (21.7)	vsurf_ID8 (13.7)
FASA_P (21.6)	VAdjMa (13.2)
PEOE_VSA_FPPOS (21.1)	AM1_dipole (13.2)
PM3_HOMO (20.5)	vsurf_DD13 (12.3)
vsurf_HB1 (20.4)	**pMRP2_c (12.2)**
PEOE_VSA_FPOL (20.2)	vsurf_Wp6 (11.9)
Q_VSA_FPOL (20.1)	PEOE_VSA+5 (11.4)
a_ICM (20)	a_nO (11.2)
SMR_VSA0 (19)	vsurf_DD23 (10.6)
AM1_HOMO (18.6)	a_don (10.2)
SlogP_VSA1 (18.6)	Num_Rings (9)
vsurf_HB3 (18.4)	PEOE_VSA-2 (8)
Q_RPC- (18)	Halogen_ratio (6.5)
PEOE_VSA-0 (17.6)	FiAB (5.4)
vsurf_ID2 (17.5)	SlogP_VSA6 (5)
Q_VSA_FPNEG (17.5)	chiral_u (4.8)
PEOE_VSA_PPOS (17.3)	**pOATP1B1_c (3.6)**
Surface_Tension (17.2)	Num_Rings_4 (3.6)
LogD(6_5) (16.8)	Rule_Of_5 (3.4)
density (16.6)	b_triple (1.1)
Kier3 (16.6)	

Lastly, similarly to what was done in Chapter 6, in order to challenge the contribution of the three physiological descriptors in the model, the model was retrained under the same conditions as the best model, with the only change being the removal of physiological features. As the best model consists of a very large ensemble (a random forest of 600 trees), it is possible that the presence of the physiological features is not advantageous and merely results from the combination of chance and the fact that the random forest is built with unpruned trees. Despite this possibility, removing the PDs yielded degradation of the model's performance (validation set MAE = 0.3003 vs 0.3044, with and without PDs, respectively). Curiously, removing three molecular descriptors of the same level of

importance as these three physiological features improved the model (validation set MAE = 0.3003 vs 0.2951, for the original and removed-descriptors model, respectively).

7.3.3. Impact of Accounting for Plasma Protein Binding

Despite the fact that PPB, as a property, is by far regarded as the most impactful physiological determinant of Vss (generally speaking, and especially compared to the remaining physiological features used in this work), including this as a feature did not produce improvement to the best model. First, putting pPPB through feature selection, alongside all other descriptors, produced a feature set that did not include pPPB (considering the full set of features provided to train the best model). As this might have resulted from overfitting to a suboptimal set of features during the pre-processing step, which is a common pitfall of genetic search, or purely due to chance (where other correlated features were selected instead), this was examined further by adding pPPB directly into the feature set provided to build the best model. This led to practically equivalent performance (validation MAE = 0.3003 vs 0.3004, with and without pPPB, respectively). However, in this alternative model pPPB showed higher feature importance than any of the three physiological descriptors used in the original best model (M5, listed in table 7.3), affecting the prediction of 20.8% of instances. This is an indication that pPPB is in fact a predictor of Vss. The full list of combinations (15 in total) of physiological descriptors for this model is provided in Table A4.2, compared to the combinations in M5.

The inability to produce an improved model does not mean PPB is not a good predictor of Vss, but rather it can just mean that this feature is very strongly correlated with other good predictive features. In this latter case, given the selection of any of the other features strongly correlated with PPB, there would be no need to include PPB in the model.

7.3.4. Benchmark Testing of the Effect of Expression-Corrected Transport and Plasma Protein Binding (PPB) as Predictors of Vss

Gombar and Hall (Gombar and Hall, 2013) used the Obach dataset to build QSAR models for Vss (as used in this work) and, given they provided individual predictions obtained in an external set, this allows a direct, benchmark assessment of the impact of adding physiological descriptors as predictors, as this is the major variation between both works. Additionally, Lombardo and Jing (Lombardo and Jing, 2016) also provide an external set of predictions, however they used a larger Vss dataset (N = 1096). Such external set allows for a benchmark testing of the value of qualitative improvement (enriching input through the

addition of physiological information) versus quantitative improvement (increasing chemical space by increasing the number of observations).

For the first benchmark testing scenario against Gombar and Hall, where the modelled dataset was the same, Table 7.4 shows that M5 shows improved performance for all performance measures used. This is particularly notable since Gombar and Hall applied two approaches that in general tend to improve predictive accuracy: they have applied physicochemical filters to assure a more tractable chemical space; and they have allocated more data for training (N=569) than in this work (N=398). Despite the expected advantage brought by these approaches, M5 still outperformed the models by Gombar and Hall, which demonstrates the value of using physiological predictors to model Vss. However, M5 did not outperform the best previous model, model 8a. On the other hand, it is important to highlight that the best model from Gombar and Hall was built using a support vector machine (SVM), which is known to be very sensitive to the set of parameters used (Mantovani et al., 2017), which usually does not happen (at least to the same extent) for RF. It could be that the SVM overfitted during parameter tuning, leading to loss in predictive performance.

Table 7.4. Comparison of predictive performance between the current best model (M5), the models by Gombar and Hall (Gombar and Hall, 2013) and the previous best model (8a), evaluated on a common external benchmark dataset (N = 30). SVM and MLR stand for support vector machine and multiple linear regression respectively.

			models in (Gombar and Hall, 2013)	
	M5 (current work)	8a (previous work)	SVM	MLR
MAE	0.209	**0.205**	0.264	0.422
GMFE	1.619	**1.604**	1.835	2.641
MFE	1.880	**1.869**	1.995	5.430

Comparing the results reported here with Lombardo and Jing's results, as mentioned earlier, M5 did not outperform their best model. However, despite the considerably smaller chemical space, M5 was still able to show comparable performance to other models as seen in Table 7.5. As seen for the comparison against Gombar and Hall, model 8a still outperformed the current best model. Both these observations might indicate that applying expression correction to transporters might not be adding value to the modelling task, however this dataset is much smaller than the test set used in this work and might not allow for the effect of such correction to be noticeable. In support of this possibility, there are particular improvements that might otherwise show the value from using expression-corrected transport data, as follows.

In their original publication, Lombardo and Hall point out three compounds in their external set which are systematically and grossly mispredicted: CG200745 (6.2 FE), dobesilic acid (4 FE) and GPX150 (5.2 FE). The predictive ability for these compounds has improved considerably using model M5 with transporter expression, as they show FE values of 3.5, 3.6 and 2.7, respectively. Rather than focusing on the numerical improvement itself, this indicates overcoming a systemic limitation found when molecular descriptors alone are used. Additionally, this has also shown an improvement comparatively to the previous model reported in Chapter 6 (model 8a), and both these observations support the importance of using transporter data corrected for tissue expression levels to aid the modelling of Vss.

Table 7.5. Comparison of predictive performance between the current best model (M5), the models by Lombardo and Jing (Lombardo and Jing, 2016) and the previous best model (8a), discussed in Chapter 6, evaluated on a common external benchmark dataset (N = 34). RF_33 and PLS_11 stand for random forest and partial least squares, respectively (number suffixes stand for the number of features used).

			Lombardo and Jing (Lombardo and Jing, 2016)		
	M5 (Current work)	8a (previous work)	RF_33	PLS_11	consensus RF_33 and PLS_11
MAE	0.3064	0.305	**0.302**	0.363	0.317
GMFE	2.025	2.017	**2.003**	2.308	2.073
MFE	2.337	**2.276**	2.300	2.970	2.510

7.3.5. Testing the Use of Physiological Predictions in Increased Chemical Space

To allow further testing of the role of PPB, PL, and expression-corrected transport as predictors of Vss, the Lombardo dataset was modelled with these additional descriptors, which were paired with either (1) this work's set of descriptors, and the current optimal modelling conditions or (2) their set of descriptors and their optimal modelling conditions (re-training), and each situation was compared with their equivalent built without PDs.

Similar to what was observed in chapter 6, in both situations (modelling with best conditions in this work or re-training using the best conditions in Lombardo et al. (Lombardo and Jing, 2016)), adding PDs improved the ability to model Vss relatively to just using molecular descriptors (see Table 7.6).

Contrarily to what was found for model M5, when the Lombardo dataset was annotated with the collection of descriptors used in this work and submitted to feature selection, PPB and pPL were selected into the final feature set. However, this model was still outperformed by M5, which is surprising given the considerably smaller chemical space of the latter

compared to the former. Comparing both "MDs & PDs" models and M5, the superiority of Lombardo's retrained model can be explained by the fact that this model has been selected as the best candidate based on the performance on this very test set, which creates a misleading outperformance.

Table 7.6. Comparison of the predictive performance of modelling the expanded Vss data with and without physiological descriptors (PPB, PL and expression-corrected transport descriptors). This is referred to as "retraining", as the modelling conditions were all kept, and merely reapplied to the larger dataset. The models were tested in the same external set (N=34) provided by Lombardo and Jing.

		Retrain with current best conditions		Retrain with original best conditions	
	M5 (current work's best model)	MDs only (baseline)	MDs & PDs	MDs only (baseline)	MDs & PDs
MAE	0.3064	0.322	**0.318**	0.300	**0.285**
GMFE	2.025	2.104	**2.081**	1.993	**1.962**
MFE	2.337	2.728	**2.572**	2.290	**2.208**
Available PDs for modelling	pBCRP1 pMRP2 pOATP1B1	n.a.	PPB pPL pPEPT1 pOATP1B1 pBCRP1 pMRP2	n.a.	all

It should be noted that both situations where Lombardo's dataset was modelled produced a larger FE error than the best model in this work (M5), for all three challenging compounds mentioned in the previous sections. In addition, the models using PDs produced smaller FE in almost all situations (except in one out of six pairwise comparisons between modelling with and without PDs).

In a recent publication (Korzekwa and Nagar, 2017) a PBPK prediction of Vss has been published using predicted PPB and a small number of physicochemical descriptors, and while its MFE is considerably lower (1.6) than any of the models in Table 7.6, which might depict this as a superior alternative to Vss prediction. However, it is important to note that all models in the work were developed and tested on the same, rather small, set of approximately 60 compounds. This is a classic example of a set of conditions that lead to high risk of overfitting and, hence, all interpretation of these results should be made conservatively. In addition this model relies on experimental physicochemical data to produce the PPB predictions. Perhaps a future option would be to replace the experimental data with in silico predictions of physicochemical descriptors, and test whether the resulting PPB predictions can be used in the current QSAR modelling scheme.

7.4. Conclusions

This study followed up on the findings in Chapter 6, where a positive impact of using physiological descriptors in the modeling of Vss was reported, and tested the hypothesis of further improvement of the predictive performance when accounting for tissue-specific transporter expression. Additionally, there was the aim of testing whether using plasma protein binding data would bring any improvement to the previous best model's predictive performance.

In this work the best produced model was able to further improve the observed performance of the best model in Chapter 6. However, in the current work's best model only transporter data (and not plasma protein binding data) was used as predictors (as opposed to the work in Chapter 6, where the best model used PL as well). The improvement was especially noticeable for the most challenging compounds, and the majority of predictions were associated with a decreased error when comparing to Chapter 6.

In the current best model, physiological descriptors showed a relatively low feature importance but, on the other hand, they have been found as top nodes in a number of trees across the random forest. This means that they are informative as predictors of Vss.

To validate the premise that expression-corrected transporter descriptors are useful features in Vd modelling, the best model in this study (M5) was compared to: (1) a model built on the same dataset as the one used here, and (2) a model built on a considerably larger dataset, using the performance on a benchmark external dataset as means of direct comparison. While M5 outperformed the model without physiological descriptors in case (1), it showed marginally worse performance than the best model without physiological descriptors in case (2). Considering that the difference in performance between M5 and the best model in (2) is very small, this is quite surprising, given that case (2) resulted from a model trained on a much wider chemical space.

Despite the apparent outperformance of the model in case (2), built without any physiological input, as observed for the test set in this work, the benchmark comparison exercise revealed that the most problematic compounds showed decreased error in their predictions.

Finally, the strongest evidence of the value of using expression-corrected transport data and PPB was the fact that the two models built from the Lombardo dataset (larger than the currently used dataset) were improved when these physiological features were used, versus when trained from molecular descriptors alone. Even though PPB did not improve the performance of model M5 when added into the feature set, this might be due to the presence

of other descriptors with strong correlations with PPB, which overshadow the useful additional contribution of PPB as a descriptor. Also, considering that there is a large proportion of compounds with missing (unknown) values for the PPB descriptor, completing missing data in the PPB descriptor with predictions derived from a highly accurate PPB predictive model could improve the quality of this descriptor when modelling Vss.

8. A Novel Applicability Domain Method: Reliability-Density Neighbourhood (RDN)

8.1. Introduction

This chapter serves the purpose of complementing the remaining chapters where QSAR models have been developed, by focusing on applicability domain characterization and its vital role in model validation. The ability to define the boundaries of the chemical space where a QSAR model can be reliably used is a necessary condition to assure the reliability of new predictions, which makes it an essential step for model validation. These boundaries correspond to the model's applicability domain (AD) and this chapter describes a novel method for AD characterization.

The theoretical goal in defining a model's AD is to identify "safe" and "unsafe" regions for prediction, which informs about reliability at various subregions across chemical space. In practice this defines the extent to which a Quantitative Structure-Activity Relationship (QSAR) model (reliably) tolerates new compounds (Eriksson et al., 2003, Carrio et al., 2014).

So far, there is no clear focus in the community for assessing whether an AD established with training data is able to successfully determine if a new prediction may be accepted or not. QSAR modellers often implement any given AD method and merely determine the portion of the external data (and its predictive accuracy) falling within the established boundaries, without any assessment of the ability of the AD boundary to differentiate between "acceptable" and "unacceptable" new predictions. Therefore, it is impossible for the user to validate and trust an arbitrary threshold. Applying a threshold and showing that, inside the region defined by that threshold, predictions have higher accuracy, as carried out in some previous work (Sahigara et al., 2012, Fjodorova et al., 2011) provides useful information, but ignores the possibility of localized inner "holes" in the chemical space where the model is unreliable.

A useful AD should relate similarly to the predictive reliability in the training set and in an external dataset. This means that both the training set and an external set should ideally have an AD profile that shows similar trends of degradation of predictive performance with increasing distance to the AD core (here the term "core" can be interpreted as the sum of one or more centroids in the AD, where predictive confidence is maximum). Ideally, a valid AD would be sufficiently robust and not affected by changes in dataset, thus allowing the

maintenance of the general AD premise by which a model's performance degrades as the queried instances get farther away from the training chemical space.

The majority of currently available AD methods usually focus on a single property of the data, for example similarity, descriptor range, density or response-range (or ensemble-range). A list of methods across categories can be found in the literature (Kaneko and Funatsu, 2014). However, several works support the need to combine different properties (such as response, density and similarity) to achieve a reliable characterization of a model's AD (Kaneko and Funatsu, 2014, Sheridan, 2012, Sahigara et al., 2013). Furthermore, most methods address data globally (e.g., location with respect to global feature span or density across global feature set), even though it is well established that the modelled data can exhibit very different properties in a local level versus the global level.

In this work a new AD method, named Reliability-Density Neighbourhood (RDN), is proposed. This method maps external predictions with regard to distance to the model space while taking into account the reliability of nearby training instances. Note that, in this context, reliability is the net effect of two distinct effects, bias and precision. As a result, RDN accounts for the variable nature of different data localities both in terms of multi-dimensional localization (as multiple dimensions are input into the distance calculation) and predictive reliability. RDN borrows features from two other previously published methods – the standard deviation (STD) method (Tetko et al., 2008) and the k-Nearest Neighbours density (dk-NN) approach (Sahigara et al., 2013).

Figure 8.1 shows a schematic depiction of the RDN AD, where density and reliability are mapped across chemical space showing densely populated and more reliable areas in darker blue, transitioning into white regions of sparse and/or unreliable data.

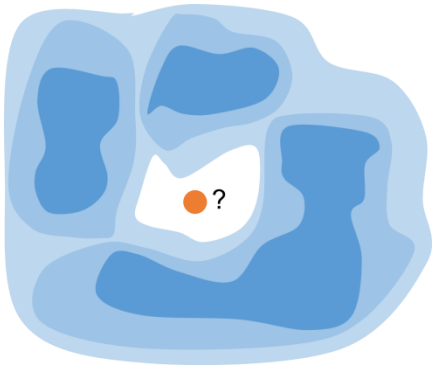


Figure 8.1. Schematic representation of how RDN explores chemical space.

This work also focuses on the role of feature selection in AD characterization, as an AD is only as explanatory as the ability of its molecular features to chemically distinguish mispredictions from correctly predicted instances. The optimization of the set of molecular descriptors used as input to compute neighbour distances is, therefore, another novel aspect of AD characterization introduced with this method.

The last novel aspect explored in this chapter is the importance of evaluating AD robustness, which was accomplished through the introduction of a new scoring scheme to evaluate the robustness and qualitative value of AD techniques.

The contents of this chapter have been published in the Journal of Cheminformatics, under the following reference: Aniceto N, Freitas AA, Bender A, Ghafourian T. A novel applicability domain technique for mapping predictive reliability across the chemical space of a QSAR: reliability-density neighbourhood. Journal of Cheminformatics, 8**,** 69.(Aniceto et al., 2016a)

8.2. The Reliability-Density Neighbourhood Algorithm

As the dk-NN approach proposed by Sahigara and colleagues (Sahigara et al., 2013) was the basis from which RDN was built, this will be described. This explanation will be built upon to transition into the RDN algorithm; its novel parameters and their contribution to the overall mechanism of this new technique will be discussed.

The dk-NN AD technique uses the k-NN principle combined with the concept of adaptive kernel techniques in Kernel Density Estimation (KDE) to detect local neighbourhoods within the data. This approach capitalizes on the notion that any given dataset can have a very different behaviour at the local level when compared to the global behaviour. In this method, the average Euclidean distance (using standardized descriptors) between each training compound and its k nearest neighbours is computed (Euclidean distance is calculated using Equation 8.1), which is used to calculate a reference value (RefVal) set at $Q_3 + 1.5 \times IQR$ (also known as the Tukey's outlier fence (Horn and Pesce, 2006)), where Q_3 is the 3rd quartile and IQR is the interquartile range calculated as the difference between the 3rd and the 1st quartiles of the list of average distances. The neighbourhood width threshold for each individual training compound (D_i) is then calculated as the average distance to all its training neighbours with distance values closer or equal to the RefVal. By establishing different local thresholds, this addresses the variation of data density across the dataset.

Given that the dk-NN AD is limited only by the degree of emptiness of the different regions occupied by the data (i.e., a sparse region will render its occupiers a smaller distance threshold, under a given established k value, as these instances will have no neighbours within the average overall distance to the k-th nearest neighbour), it would be logical to tailor each different neighbourhood (i.e. coverage width around each training instance) according to their reliability. Here, reliability is a function of both bias and precision as explained below.

$$\text{Euclidean Distance}(a, b) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \qquad \text{Eq. 8.1}$$

Following the theoretical principle that, for more reliable predictions, an ensemble (set) of models, M, will have a high degree of agreement and consequently a smaller STD (standard deviation) value, one would expect that regions where a clear, smooth structure-activity relationship is found would generate more robust predictions that are less susceptible to changes in the learning task (i.e., changes in the data partition within the ensemble). Alternatively, regions with a less stable landscape will rely greatly on the data partition used, thus generating larger differences between different models (Sahlin, 2013). However, as STD values only measure the level of precision, the rate of agreement between the set of predictions and the real responses needs to be used to overcome cases of systematic bias towards an incorrect classification. More precisely, a systematic bias occurs when most predictions are close to each other, but all are wrong, as represented by the point filled in black in Figure 8.2. These predictions would be captured by the algorithm as high-reliability predictions if only an STD correction was used. Therefore, the combination of bias and precision is an appropriate correction factor for reliability, W_i.

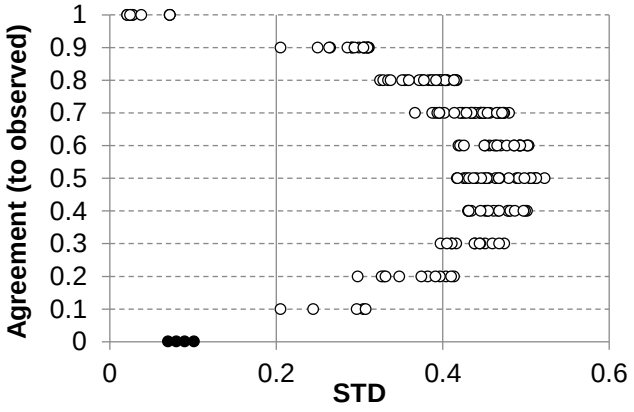


Figure 8.2. Relationship between agreement and standard deviation across the members of an ensemble in the P-gp validation dataset. In this case STD translates into accordance among a set of predictions (i.e. precision), whereas Agreement refers to the level of bias in that set of predictions.

Taking this notion into account, the RDN AD method herein proposed was created by introducing a weighting term to the dk-NN algorithm, as defined in Equation 8.2, which measures the reliability associated to each training instance.

$$W_i = \left(1 - \underbrace{\sqrt{\frac{\sum_{m=1}^{M}\left(\widehat{y_{i,m}} - \bar{y}_i\right)^2}{M-1}}}_{\text{STD}}\right) \times \underbrace{\frac{\left|Y_i \cap \widehat{Y}_i\right|}{M}}_{\text{agreement}} \qquad \text{(Eq. 8.2)}$$

The first term (1-STD) measures precision and the second term (agreement) measures bias. In this equation, the weighting factor $\widehat{y_{i,m}}$ is the predicted class probability for compound i, output by model m, among M models in the ensemble; $\bar{y}_i$ is the average predicted class probability by the ensemble model; Y_i is the experimental response; and $\widehat{Y}_i$ is the prediction output by the QSAR model. As STD and agreement take values from 0 to 1, W_i will also take this range of values.

For each training instance i, W_i will be multiplied to the respective threshold distance D_i, calculated as previously explained. As STD is the deviation between an ensemble of predictions, 1 – STD is the precision rate. A high precision will translate into a high 1 – STD value which will, in turn, contribute to a large W_i, and consequently to a small reduction of D_i. As for the agreement term, increasing values translate into a decreasing level of bias. As such, a large agreement will entail a small penalization to D_i. To illustrate the use of W_i, the space (neighbourhood) covered by a given training point will be penalized proportionally to its degree of unreliability, i.e., for STD=70% and agreement = 35%, a reliability of 10.5% is obtained, which leads to a very large 89.5% reduction of coverage attributed to its training instance. The effect of correcting neighbourhood distances for their reliability is demonstrated in Figure 8.3. The complete flow of the described RDN algorithm is summarized in Scheme 8.1.

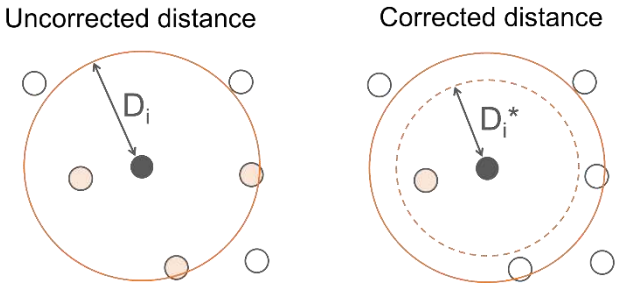
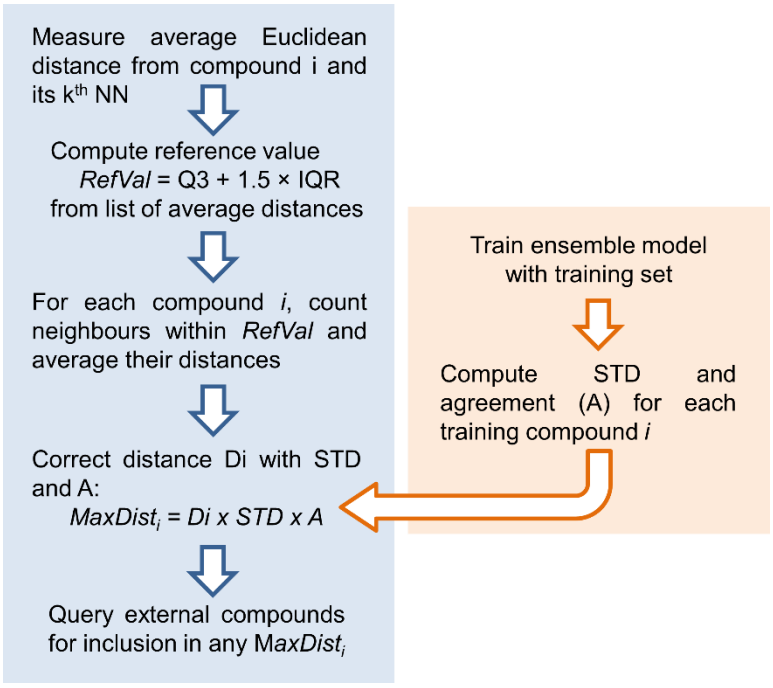


Figure 8.3. Scheme of the reliability correction of the distance D_i attributed to training compound i. The sphere's radius, Di, will be decreased proportionally to the reliability of compound i. For example, if (1-STD)×agreement is 80%, D_i will be reduced by 20% of its initial value, which means that the 2

of the initial 3 external instances that were covered by compound i will end up outside the neighbourhood coverage area associated with this training compound.

The success of addressing local bias and precision, as well as local distance to training has been demonstrated by Sheridan (Sheridan, 2012); however they have sorted the data into several bins, which renders comparative analysis and the implementation of the AD rather difficult. A continuous performance characterization should allow the localization of gaps in the data/model's chemical space in a more user-friendly way.



Scheme 8.1. Pseudo-algorithm of the Reliability-Density neighbourhood (RDN) applicability domain technique.

As the obtained individual thresholds associated with each training instance depend on the Euclidean distance between compounds, which in turn depends on the descriptors used, this chapter proposes the approach of pairing this AD technique with a feature selection technique applied in a preprocessing step (before running the classification algorithm). ReliefF was chosen, originally proposed by Kononenko et al. (Kononenko et al., 1996), as this algorithm searches for a feature set that maximizes the separation of classes in the response variable within local neighbourhoods (Spolaôr et al., 2013). ReliefF has been shown to detect relevant features even in very crowded (feature-wise) datasets, whilst being resilient to noise (Bolón-Canedo et al., 2013, Robnik-Šikonja and Kononenko, 2003). ReliefF is particularly well suited for AD definition due to three paramount properties: a) it evaluates descriptors separately and solely on their ability to separate classes; b) it takes into account the local neighbourhoods when evaluating each feature; c) identifies useless/irrelevant features that would only contribute with noise (Hall and Holmes, 2003). Regarding the first property, while ReliefF allows the selection of highly correlated features,

its performance is unaffected by the existence of correlation itself (Kantardzic, 2011); contrarily to QSAR modelling, this is expectedly a desirable feature for a successful AD as highly correlated features turn out to be complementary in chemical space coverage. This is further explored in the Results and Discussion section of this chapter (section 8.4).

Considering that a QSAR model is focused on distinguishing between two different responses, and its AD is focused on discriminating between correct and incorrect predictions, it is expected that the molecular descriptors that are best suited for the former will not necessarily be the most appropriate for the latter, as previously suggested (Sheridan, 2012). In fact, Sheridan and colleagues (Sheridan et al., 2004) have shown that descriptors used to define the model's boundary do not have to coincide with the descriptors used to build that same model. Furthermore, note that an AD technique which does not rely on the features used by the QSAR model allows comparable implementation in both the so-called transparent methods (e.g. decision trees) and "black box" methods (e.g., artificial neural networks). Thus, the herein proposed AD method is paired with the ReliefF routine for feature selection.

8.3. Methods

8.3.1. Building of the QSAR model

To evaluate the performance of the currently proposed AD, the QSAR model previously built with the P-Glycoprotein (P-gp) dataset (extracted from the multi-label ABC efflux dataset) was used (see Chapter 3 for further details regarding data retrieval and preparation). This is a classification dataset annotated with substrates and non-substrates. Essential details will be reiterated here, as full details on the pre-processing feature selection and modelling procedures can be found in Chapter 4.

Recall that a decision tree was trained using 60% of data (training set), optimized using 20% of the data (internal validation set), and tested on the remaining 20% (test set). Pre-processing feature selection was performed prior to training, by submitting the training set to the five feature selection methods described in Chapter 3 (ReliefF, GS, GA, RF-GS, C4.5-GA). The C4.5-GA wrapper method was selected to build the final decision tree model as it generated the feature set associated with the highest validation performance. The resulting decision tree was used to produce class predictions, which were later used to evaluate AD performance. Note that the feature selection task undertaken within the model building

process (described under this subsection) must not be mistaken for the feature selection role in establishing AD characterization. These two are separate and independent tasks.

8.3.2. Feature Selection in AD characterization

To establish an optimal feature set utilized in the RDN algorithm, more specifically in the calculation of the Euclidean distance between the compounds in the P-gp dataset, different thresholds of feature ranking using ReliefF were applied, namely the top 20, 50, 100 and 200 features, as well as the entire feature set of 334 molecular descriptors. This led to 5 feature sets that were tested in the original dk-NN algorithm. For comparison, the C4.5-GA features used to train the QSAR model were also used, as it is a common practice to use the model's features to describe the AD. RDN was not used to assess the effect of the descriptor sets as this would introduce additional noise to the system (due to different variables in play) and could confound the comparison between feature sets. As dk-NN takes into account solely the Euclidean Distances between compounds, this allows a more straightforward observation of the effect of the feature set. Furthermore, a selection of the best feature set candidate(s) in RDN would increase the risk for parameter overfitting. At the end of this stage the two best candidates were selected for further testing with RDN.

8.3.3. Consensus Standard Deviation (STD) Applicability Domain

Even though the STD measure was embedded in the RDN algorithm as part of the correction factor, this is a standalone AD method that has obtained very good performance in sorting predictions according to their reliability. As a result, STD was used as the gold standard method against which RDN was compared (Sushko et al., 2014, Dragos et al., 2009, Tetko et al., 2008, Sushko et al., 2010a). Note, however, that the results of dk-NN and KDE methods will also be reported for comparison (these methods are explained further below).

For the implementation of the STD method, a 10-fold C4.5 bootstrap routine was performed following the procedure in the methodology Chapter 3. This resulted in 10 decision trees which were used solely to produce reliability estimates in the form of overall deviation among the 10 sets of predictions, while class predictions were taken from the decision tree model reported in Chapter 4. The STD value was calculated for each compound according to Section 3.6.

Contrarily to the QSAR model whose output is ultimately qualitative (an instance is assigned to the class of highest probability), the actual value of the probability was used towards the quantification of reliability. Consequently, probability calibration by Laplace smoothing (for a detailed outlining see (Chawla, 2006)) has been used during the training of the ensemble model. Laplace smoothing compensates for the small number of compounds in a tree node, thus preventing overly optimistic probabilities at very small nodes.

8.3.4. Reliability-Density Neighbourhood Applicability Domain

The RDN AD was implemented as described in the Algorithm section 8.2, being run iteratively at increasing k values, ranging from 1 to 65 (weighted) nearest neighbours (NN), which corresponds to approximately 100% coverage of the data (as obtained empirically). This allows to scan the chemical space from denser areas to sparser areas. Preliminary results showed that using the distance step size to the first NN directly was not ideal, as the AD RefVal led to a too wide AD (with more than 50% of data falling within the nearest 2-3 neighbours region). This is because this region is more densely populated, thus being highly sensitive to even small increases in the distance threshold (see Figure 8.4). Therefore, it is necessary to make sure that the initial neighbourhood thresholds increase slowly. Then, as the AD boundaries get larger, it is affordable to have larger distance increases at each step. To this end, the RDN algorithm was run at a third of the determined neighbourhood distance from k = 1 to 30, then half of the neighbourhood distance was used for k = 31 to 40, and finally for k values > 40 the distance was used directly as computed. However, this setting can be tailored by the user, and different distance step sizes can be used to obtain different levels of detail in the plots of accuracy vs percentage of data in the AD. As exemplified in Figure 8.4, initially applying smaller increments in distance thresholds (right-hand side) allows a slower inclusion of data into the AD, which consequently improves sensitivity at the inner core of the model.

As originally implemented in the dk-NN algorithm, a query must fall within the neighbourhood threshold of at least one training instance in order to be considered inside the AD. This prompted the assessment of the impact that the number of required training neighbours has on the overall performance of the AD. To do so, the algorithm was tested with different minimum required k values which offer coverage to new instances, ranging from 2 to 30.

For the calculation of the RDN AD profile, W_i (Equation 8.2) is calculated for each training instance to correct their neighbourhood radius distance according to their level of precision and bias. For the P-gp model, STD was calculated from the deviation observed across a

10-fold bagged decision tree ensemble, as shown in section 3.6. Regarding the values of agreement, these were calculated by determining the frequency of predictions in the ensemble which were correct (i.e. matching the observed class).

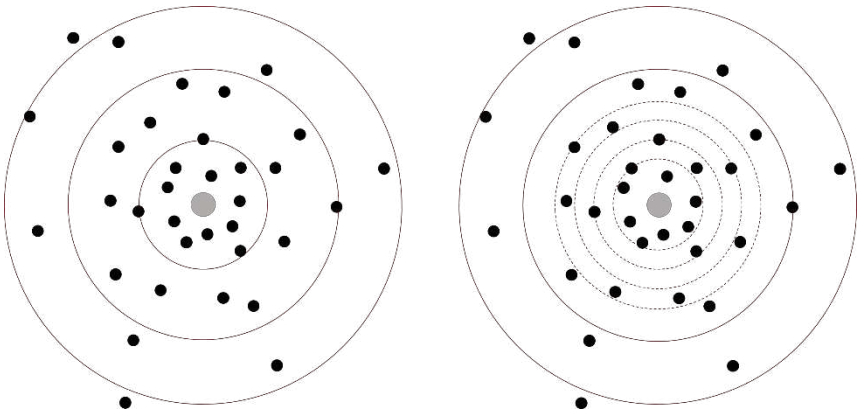


Figure 8.4. Schematic representation of the difference between the RDN algorithm without (left) and with (right) distance step adaptation. The grey point represents a training instance, and the black points depict external instances scattered across a 2D projection of the 20 molecular feature matrix. Smaller increases in radius around the training instance in grey increase sensitivity in measured accuracy across the AD landscape.

8.3.5. Comparison with dk-NN and KDE AD Methods

For a comparison, STD and dk-NN methods have been implemented as they both are integrated in the RDN algorithm. The implementation of both was done as described earlier. Additionally, Kernel Density Estimation (KDE) has been used for its specific features which address data from a different perspective. Similarly to k-NN, KDE addresses data density, however the former focuses on local neighbourhoods, whereas the latter addresses overall data density across descriptor space. Since RDN accounts for both density and predictive reliability, it is worth evaluating both density in chemical space (both locally and globally) and response distribution separately. KDE was computed using KernelDensity within the sklearn python module, in which a Gaussian kernel was used and the bandwidth was selected from an online platform (http://176.32.89.45/~hideaki/res/kernel.html) of bandwidth optimization created by Shimazaki and Shinomoto (Shimazaki and Shinomoto, 2010). The implementation of KDE followed the procedure outlined elsewhere (Jaworska et al., 2005). The density distribution model was established from the first principal component obtained from the training set, and the validation and test sets were matched against it to test the hypothesis of density being correlated with predictive accuracy (i.e., accuracy decreases with decreasing density).

Furthermore, as the P-gp model was built using a decision tree learner, it is worth monitoring misprediction occurrences with respect to the chemical span in the decision tree's branches. This analysis aimed at identifying any trends within the decision tree's chemical space subpartitions.

8.3.6. Quantitative Comparison Between AD Methods

In order to establish which AD method yields the best performance, a scoring function was proposed which aims for a quantitative, objective comparison between methods. This scoring function evaluates two features: (1) robustness, by measuring the similarity between the AD profiles of two external datasets, and (2) proximity to a smooth descending AD profile (accuracy vs the AD-produced measure of prediction confidence).

This scoring function is meant for the scoring of continuous ADs, not being suited for in-out binary type approaches. As any AD method is only reliable if it is robust when submitted to different subsets of the same dataset, this AD scoring function will quantify the ability of an AD to produce the same outcome in two different external datasets Y and Z. In an ideal scenario, where the AD of a model is mapped in a robust manner across the training data, Y and Z would yield two perfectly matching curves of accuracy vs distance-to-model (DTM). This indicates that the model's reliability readout (i.e., a trend between predictive performance and the AD measure) is not being affected by the specific dataset being evaluated, but instead the AD is robust enough to describe the predictive reliability across the data. Additionally, in the curves for both datasets Y and Z, the accuracy inside the AD boundaries should decrease steadily as a function of DTM, as it is theoretically expected that a model's performance will degrade as the distance to the training space increases. Equation 8.3 quantifies both aspects and produces a final score.

$$\text{AD score} = \frac{1}{F_{\text{added},[1;P]}} \sum_{i=2}^{P} WP_i \times |y_i - z_i| + WP_i \qquad \text{(Eq. 8.3)}$$

In this AD scoring function, (y_i - z_i) quantifies the accuracy difference at each AD distance i, and WP_i stands for weighted slope mismatch penalty at distance *i*, which measures the mismatch between the curves' directions at each distance interval. This will cover the entire curve of measured ACC vs AD measure across all points, P. A weighted measure was used for the slope mismatch explained below. More specifically, as each distance point is associated with a given amount of newly added instances (N_{added}) into the AD, the slope

mismatch penalty is weighted according to how many instances have been added at a given distance interval (Equation 8.4).

$$\mathrm{WP_i} = \mathrm{SMP_{[i,i-1]}} \times \frac{\mathrm{N_{added,i}(y+z)}}{\mathrm{N_{total}(y+z)}} \quad \text{(Eq. 8.4)}$$

As the AD is expanded (DTM is being increased), the directions of the two curves are monitored using a term that penalizes slope mismatch between the curves, the Slope Mismatch Penalty (SMP). A qualitative penalty scheme that differentiates the various types of mismatch was set, described as follows (see Figure 8.5):

The slope, m, of any segment in an AD curve (between distances i and i-1) can be m = 0, m > 0 or m < 0. Considering the requirement that accuracy should decrease with respect to distance-to-model, it is reasonable to consider m<0 as the desirable case, m=0 as less desirable and m>0 as the least desirable case. As such, a multiplicative penalty of 1 (i.e. no penalty) has been attributed to a negative slope and the penalty doubles consecutively for a null slope and a positive slope (i.e., 2 and 4, respectively). This set of penalties was optimized to allow a correct scoring of a positive control (a visibly highly similar pair of curves) and negative control (a visibly highly dissimilar pair of curves), such that the former is the least penalized and the latter is the most penalized scenario. To compare two corresponding pair-wise segments, each segment on both curves is attributed a penalty according to its individual slope. The resulting product of the individual penalties of those two equivalent segments between i and $i\text{-}1$ of the curve corresponds to SMP_i. The various possible scenarios are exemplified in Figure 8.5, where they are organized from the most desirable to the least desirable (from A to F, respectively).

The weighting of SMP by the amount of data points that are added to the applicability domain with each step of increased distance-to-model allows accounting for different local densities, which is necessary considering that a shift in the slope direction is more significant if it is caused by the addition of, for example, 50 new data points than by 2. As the scoring function is comparing each pair of corresponding points in both Y and Z curves, the numbers of instances under such pair of points are added together and divided by the total number of instances of both, to allow comparison between AD techniques that produce a different amount of distance-to-model points.

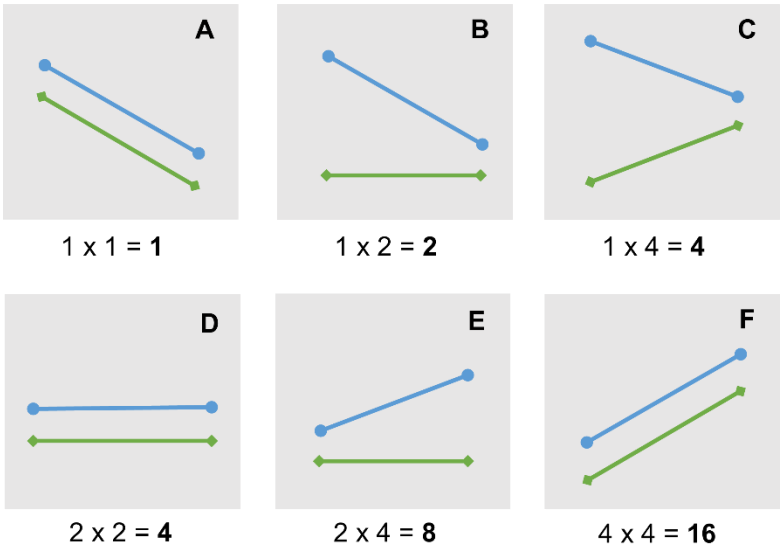


Figure 8.5. Representation of the different possible Slope Mismatch Penalties, organized from the most desirable (ideal) scenario in A to the least desirable scenario in F.

In addition, the absolute difference of accuracy ($|y_i - z_i|$) under the same distance-to-model value (X-axis) is also included in the AD scoring function. This corresponds to the underlying concept of the Fréchet distance commonly used to measure curve similarity (Efrat et al., 2006). However, this is not a decisive aspect, since a shift in absolute accuracy values will not have any impact in the decision of accepting or rejecting any given prediction, as long as the AD curves match in shape (i.e., the highest accuracy occurs at the same region for both curves). As a result, this is included with the sole purpose of allowing to differentiate between two pairs of curves where, in each pair, both curves have exactly the same shape within the pair, but one pair shows larger deviation of absolute accuracy values. To prevent this parameter from having a large impact on the total score (which would be inappropriate), it was added as coefficient of WP, as depicted in Equation 8.3.

Lastly, as different AD techniques cover a different amount of data with their first iteration, which can be regarded as the AD's core, it is desirable to differentiate between AD techniques according to their resolution at the model's core. It is more useful to cover 5% of the total data with the first iteration than 50% of the data, as the user has no information regarding the accuracy vs distance relationship across that portion of the data. As a result, the final sum across all distances i is divided by the fraction of covered data from the first iteration to the last (F_{added}); as this value approaches 1, the resolution at the model's core increases, and the final sum is increasingly less inflated.

8.3.7. Testing on Benchmark Datasets

To exclude the possibility of an exceptional performance under the P-gp dataset, two benchmark classification datasets were tested: the Ames mutagenicity dataset ("Ames levenberg" model entry, referred to as "Ames" from now on) and the CYP450 inhibition dataset ("CYP450 modulation e-state" model entry, referred to as "CYP450" from now on). To avoid any additional bias, the datasets were previously modelled (Sushko et al., 2010b) and the predictions were used as provided at the OChem QSAR modelling repository (https://ochem.eu/home/show.do). To allow testing the robustness of the AD profile, the validation datasets retrieved from OChem were split into two. Therefore, in this work, AD was evaluated in the P-gp model using the validation and TE, and the AD of the two models of benchmark datasets was assessed by splitting the provided external dataset into two sets of data. The Ames dataset comprised a training set of 4358 compounds, and two external sets of 1089 and 1090 compounds. The CYP450 dataset comprised 3743 training compounds, and 1870 compounds in each of the external test sets.

To maximize direct comparability, the source of the feature set used in every AD technique implemented for each dataset was kept fixed. As the purpose of this study is to validate the observed profile with the P-gp model, upon which the RDN technique was optimized, the feature selection procedure used in this case (i.e., top 20 features selected by ReliefF) was applied to the benchmark datasets. This potentially avoids background confounding that might perturb the effect of the AD method being applied to a given dataset.

For the calculation of the RDN AD for the two benchmark models, STD was used as provided in the OChem platform (calculated using the same method as described in this paper). As the output probabilities of each model of the ensemble were not available for the benchmark models, the agreement values were calculated from the inverse of the difference between average predicted probability and the observed value (so, an average predicted ensemble probability of 0.23 for an observed class value of 0 equates to $1 - |0 - 0.23| = 0.77$ agreement). Even though this is more skewed than the frequency of correct predictions, it still represents the majority vote (or the overall predictive trend), to some extent. In fact this is a more conservative way to calculate the agreement, since larger agreement values are only achieved when the majority of the predictions also have a value close to the expected class, and it is no longer sufficient that the majority is merely beyond (above or below) a threshold of P=0.5.

Note that, to allow a closer analysis of the rate at which data is being included at each iteration of each AD method, all AD profiles will be presented in the form of Accuracy as a function of amount of data included into the AD. As different AD techniques often generate

different types of threshold values (number of neighbours, standard deviation, and density percentile), this standardization also allows a simpler and more intuitive visual analysis of the readouts. However, attention must be paid to the fact that the actual establishment and use of each technique relies solely on the output measures. So, two profiles for the same technique applied to the same dataset under different parameters (e.g., a different set of features) might generate a percentage of 15% and 70% of included data, respectively, within their first iteration. If this first iteration is measuring the average distance to the first nearest neighbour, both cases will compute this distance differently (due to the use of different parameters), which will in turn generate a larger or smaller inclusion of data.

8.4. Results and Discussion

8.4.1. The Role of Feature Selection in Establishing the RDN Method's AD for the P-gp Dataset

Firstly, the original dk-NN was implemented on the validation set using different sets of features to assess the impact of different sizes of the feature set. Figure 8.6 shows very different AD curves for different features used. Interestingly, the feature set leading to the best validation performance in the P-gp model development (Aniceto et al., 2016b), namely C4.5-GA-derived features, revealed to be far from acceptable for AD characterization using this technique, as the smallest achievable region around the AD core includes almost the entire dataset (91.8% coverage) and it shows an accuracy of 0.685 – which is below the baseline accuracy of the global validation set at 0.691. This is in line with the theoretical expectation that the training of the QSAR model and the calculation of the AD are two different tasks, as already explained earlier in the chapter.

The AD profiles built from all features and from the ReliefF top 20 features were the best ones, showing signs of decreasing degradation as the distance to the model's core increases. As this indicates the possible ability of these two feature sets to locate higher quality predictions at the model's core, both feature sets, namely the ReliefF top 20 features and all features, were tried in the RDN AD development as well as the model's feature sets, C4.5-GA, for comparison. Figure 8.7 shows that by using the ReliefF top 20 features a better resolution is achieved at the model's core. More precisely, using all features leads to the inclusion of ~80% of the external data at the first iteration, while using the ReliefF top 20 features, only ~62% of the data is included in the first iteration. Also, both the ReliefF top 20 and C4.5-GA curves show a statistically significant difference (Wilcoxon paired signed

rank test, p = 0.0270, carried at a 95% confidence level after a failed Shapiro-Wilk normality test).

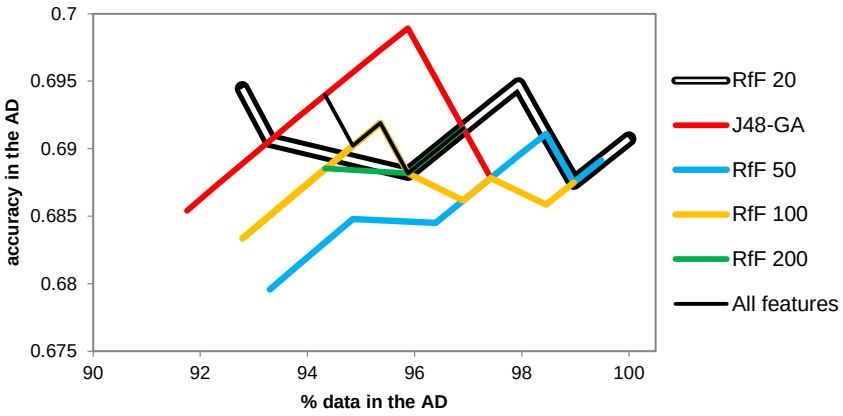


Figure 8.6. Comparison of different feature sets used in the dk-NN AD by Sahigara et al (Sahigara et al., 2013), applied to the P-gp validation set. The baseline value (i.e., accuracy corresponding to 100% data inclusion) for the IV set is 0.6907.

In addition, the RDN AD developed by using the ReliefF top 20 features shows a visible decline in accuracy as the distance to the model's core is increased (by addition of new data). This shows an improvement when compared with dk-NN AD developed by this same set of molecular descriptors (compare Figures 8.6 and 8.7). This means that penalising the distance thresholds attributed to each training instance according to their reliability (measured in STD and agreement) is useful towards mapping an AD with a higher quality core.

Results show that neither of the feature options commonly used in AD development – i.e. the model's descriptors or all available descriptors (Tropsha and Golbraikh, 2010, Dragos et al., 2009) – were appropriate for this dataset. The lack of ability to differentiate high reliability regions and low reliability regions across the chemical space when using all features is probably a sign of an overwhelming amount of noise that prevents the algorithm from taking advantage of meaningful variables. This goes against expert recommendation that all available features should be used (Tropsha and Golbraikh, 2010). Even if these observations do not necessarily apply to each and every QSAR problem, they should at least raise awareness to the fact that a feature selection routine should be carried within the task of AD characterization.

It would be theoretically expected that C4.5-GA would lead to a better AD characterization as it yielded a better learning performance which, in practice, means that it generated a

decision tree better able to differentiate the two classes. However, the herein reported results show that ReliefF was visibly better able to generate more informative features with respect to misprediction-correct prediction separation (Figure 8.7). Considering that classification errors happen by lack of ability to differentiate the two classes at certain regions of the chemical space, it is possible that features that directly address class differentiation are more explanatory in these problematic locations of the structure-activity landscape.

The reason why ReliefF outperforms C4.5-GA in this particular task might be because it selects relevant features even if they are highly correlated to other highly ranked features (Kantardzic, 2011, Hall and Holmes, 2003). This is possibly advantageous when defining the AD as two features might be highly correlated but still necessary to provide chemical coverage at specific locations of the data, which can be interpreted as feature cooperation – recall that feature dependencies can potentially hold information that an isolated feature cannot represent, as exemplified by Dragos et al. (Dragos et al., 2009) (highly correlated hydrogen bond donor capacity and (positive) charge provide potentially essential information when combined). This ability to capture local idiosyncrasies and to uncover informative label interactions are some of the strongest characteristics of ReliefF (Bolón-Canedo et al., 2015, Hall and Holmes, 2003, Bolón-Canedo et al., 2013), and it has been recommended as useful when the task can take advantage of strong feature interactions (Hall and Holmes, 2003).

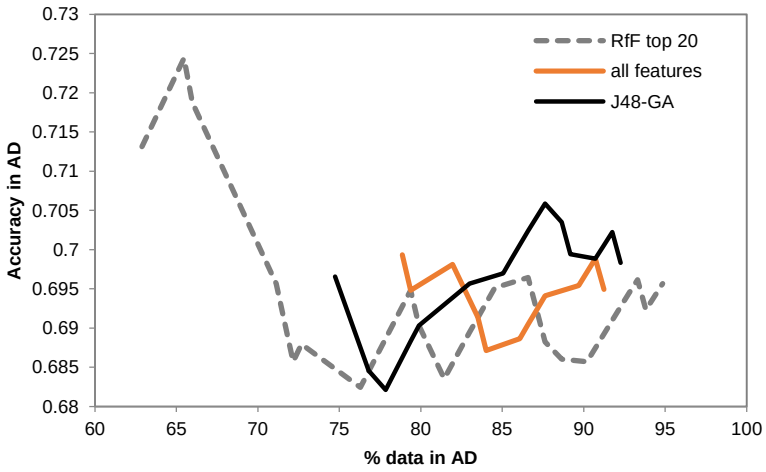


Figure 8.7. Comparison between RDN applied to the P-gp validation dataset using the ReliefF top 20 features, all features or the features selected by C4.5-GA. Note that this implementation of RDN corresponds to using the distances directly from the k-average nearest neighbour (i.e., the distance shrinking to 1/3 and 1/2 has not been applied yet at this point, as explained later in the discussion).

In addition, using a wrapper means the bias of the C4.5-GA feature selection algorithm interacts with the bias of the C4.5 learning algorithm (Tang et al., 2014, Liu et al., 2010). Tetko et al. (Tetko et al., 2008) reported that using the descriptors previously used to train the model does not lead to a better AD. This is in line with the observation that the features used for the modelling did not yield the best AD. Given that ReliefF generated high quality AD for the benchmark dataset (discussed below), this study proposes that this technique is, in principle, particularly well-suited for AD mapping.

8.4.2. Implementation of the RDN-AD Using the ReliefF top 20 Feature Set

Even though using the ReliefF top 20 features yielded a visible improvement in the AD quality, Figure 8.7 shows that, at this point, the RDN technique is still insufficient in mapping the reliability close to the model's core, as taking into account the region up to the average 1st nearest neighbour satisfies more than 60% of validation data. Hence it can be deduced that the supposed inner-most region of the AD is far too large to be able to sort predictions for their reliability. This led to the implementation of three different distance steps as the neighbourhoods are increased (as described earlier in this chapter, Section 8.3). It was hypothesized that, as regions closer to the AD's core are expected to have more data, this area requires smaller steps for increasing distance, and as distances to the training data get larger the distance increment step can also increase. Applying this modification in distance step size did in fact bring a marked improvement in the quality of the AD core, as depicted in Figure 8.8 by the higher accuracy value at the first iterations of ReliefF top 20. As explained before, recall that the percentage of included data is a mere result of an underlying distance-to-model threshold measure. As a result, the first point in both profiles corresponds to the same iteration (which in this case is the respective average distance to the first nearest neighbour). Additionally to this, ReliefF top 20 also yielded better resolution at the AD's core (a smaller portion of data included at the first iteration, which allows a more gradual monitoring of quality across chemical space).

Furthermore, there is a marked difference between the initial dk-NN-derived profiles and the final RDN profiles (Figure 8.8, A vs B). Considering that the dk-NN method can be regarded as the backbone of the RDN technique, this marked improvement in the ability to sort external set predictions according to their reliability is attributed to taking into account the local bias and precision (the correction factors), as well as allowing a slower increase of the AD span (i.e. slower scanning from the core to the outer regions of chemical space).

Figure 8.8 B shows that even though the accuracy vs size of the AD is not a smooth profile, it shows a very similar trend between the two external sets (validation and test sets). There

is a main accuracy drop in the RDN AD at around 15% of data in the AD, which corresponds to a specific Euclidean distance from every training instance. So, it is probable that the chemical space corresponding to instances that fall around this distance is problematic. As a consequence, more importantly than having perfectly smooth profiles of degradation with respect to distance to the model, it is a priority that the established AD profile (in this case through the validation set) is able to correctly characterize how new data will behave, in a robust manner, across chemical space. One should remember that other issues of the model are being brought along with any AD assessment, i.e., activity cliffs, experimental errors in the response variable, and specific shortcomings of the machine learning task undertaken (e.g. overfitting).

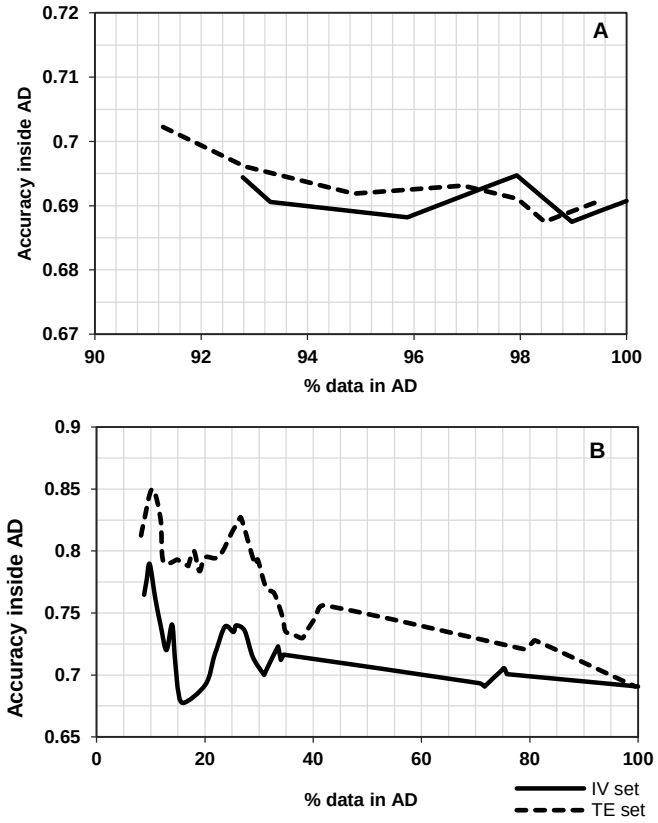


Figure 8.8. Comparison between dk-NN (A) and RDN (B) ADs, both computed using the top 20 ReliefF selected features applied to the P-gp dataset. RDN was implemented with different distance increase steps as explained in the Section 8.3.

Note that the percentage of inclusion and accuracy are cumulative. So, as the model space is being further explored, whenever an unreliable region is reached the detrimental effect of poor accuracy associated with compounds in this region will be propagated to the following regions, and their accuracy values will be deteriorated. This means that, when a low quality patch is found around the area corresponding to 15% of included data, this will decrease

the accuracy at the following regions, which means that quality at the location of 23% inclusion would actually be higher than the observed 74%.

In an attempt to establish the cause for the abrupt decline observed at the beginning of the AD curve in Figure 8.8 B, the compounds entering the AD around 15% of included data were analysed. The descending part of the curve that precedes this point corresponds to 4 compounds being added through 4 distance steps (4 iterations of the algorithm), which in itself indicates this is a sparse region of the model. As a consequence, it is understandable that 3 of those 4 instances are mispredicted, given the theoretical link between data density and predictive confidence. It would be very difficult for the model to properly establish any link between structure and activity dependence with such scarcity of information on both aspects.

Looking into the absolute maximum (model's core) of the AD, it was observed that the 18 molecules covered at this point are generally very dissimilar (similarity matrix in Appendix V, Figure A5.1), showing a 0.1137 median Tanimoto coefficient of ECFP4 fingerprints, which spanned between 0.029 and 0.71. This rules out the assumption that the model's core corresponds to a cluster of data – which would render this AD very limited for new data; instead the model's core is spread across chemical space, into various smaller sub-portions of the core.

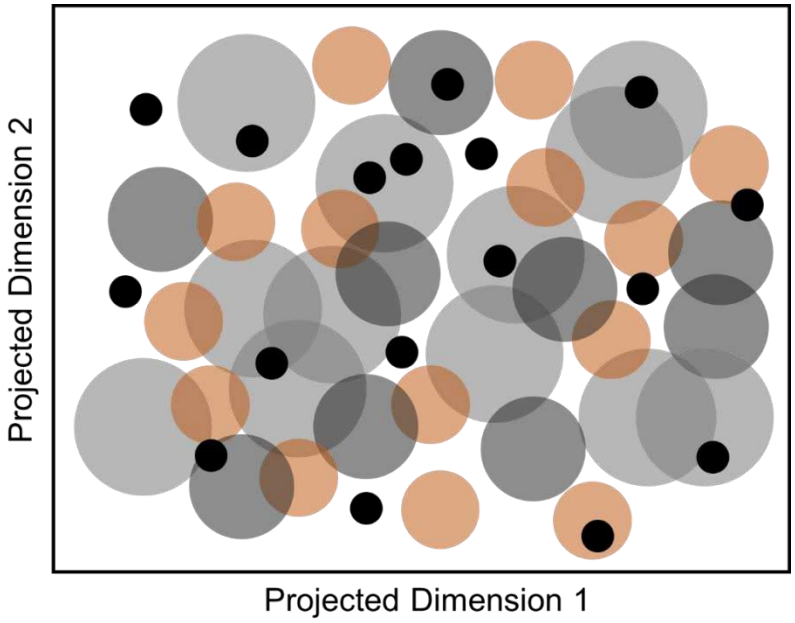


Figure 8.9. Visual representation of the RDN AD across two projected dimensions of the input set of molecular descriptors. Larger (light gray) circles are established from training instances with higher density and/or higher reliability (small bias and large precision), and as circles decrease in size (dark gray, and orange) this indicates less dense/reliable regions of training space. External test predictions (black) are placed onto chemical space and if covered by any of the training circles they are deemed as being within the AD, for the established distance threshold.

Figure 8.9 shows a graphical depiction of the neighbourhood circles around the training space, and how the external set scatters with respect to it.

8.4.3. Comparison between RDN and STD ADs

Ensemble standard deviation (STD) and STD-related methods are arguably some of the most successful AD techniques in the literature (Sushko et al., 2014) (see comparative studies in (Tetko et al., 2008, Dragos et al., 2009, Sushko et al., 2010a)). As a result STD was set as the “gold standard” comparator, and comparisons will be made with respect to test set performance, and degree of matching between validation set and test set.

Figure 8.10 shows the STD AD profile for training, validation and test sets as a plot of accuracy vs the standard deviation between the ensemble predictions. Firstly it is important to note how misleading it is to use the training set to define the AD, as commonly done by QSAR practitioners. As clearly shown in Figure 8.10, the training set gives an overly optimistic reliability profile across STD, which stems from the natural tendency for overfitting, and also possibly due to the systematic bias for the external sets. In this scenario, it is preferable to have a conservative reliability profile given by the validation set, which is what was done with the RDN AD above.

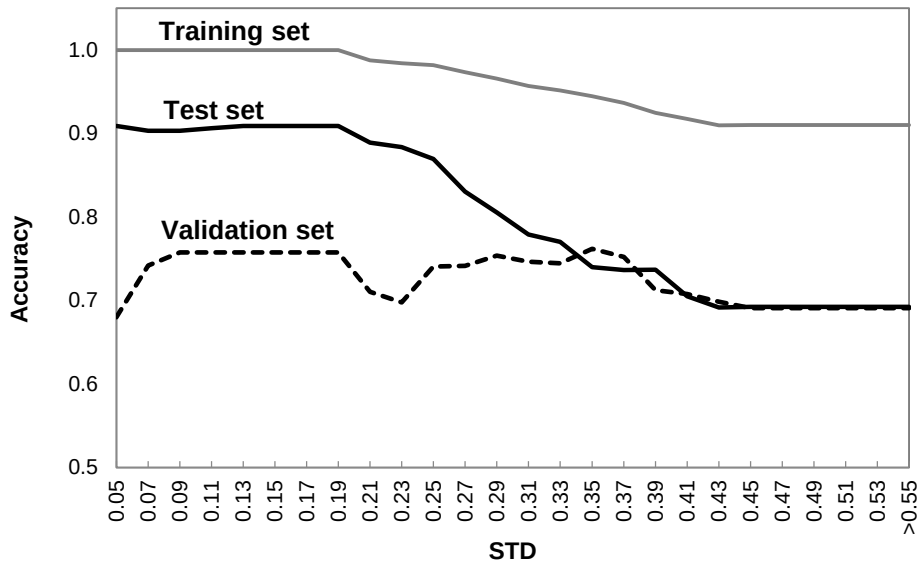


Figure 8.10. Accuracy across STD tiers for the different P-gp datasets.

Even though STD shows a very smooth profile on the test set, this does not mean that STD outperforms RDN, as the addition of new compounds is based on the standard deviations of predictions by various ensemble models, which is a more supervised procedure than

RDN (Figure 8.8 B) where compounds were being added based on the corrected distance to training space. In addition, Figure 8.10 shows that there is a marked difference between test set and validation set accuracy profiles across the AD, which renders this technique unpredictable with new data. This difference stems from the fact that low STD does not necessarily mean high quality of prediction, and it merely translates into high precision of the machine learning task – the lack of sensitivity to bias is the main flaw of this method, which is addressed in the newly proposed RDN method through the addition of the weighting term W_i (which accounts for both precision and bias). Therefore, different datasets suffer, to different extents, from systemic bias when training a QSAR model. This phenomenon can be demonstrated by the notable impact that accounting for bias (by using the agreement measure) has in both profile smoothness and inner-core quality (Figure 8.11). If agreement is taken into account, situations of high precision-high bias (affecting the quality of the STD AD) are overcome for the validation set. This observation further supports the use of both precision and bias measures as correction factors in the RDN algorithm.

The RDN outperforms the STD method as the former shows similarity between the accuracy profiles of both P-gp external datasets (validation an test sets) as well as showing similar accuracy levels for these two sets, which is an evidence that this AD method appropriately addresses data locality, while the STD AD method shows high discrepancy between the two sets (as depicted in Figure 8.10).

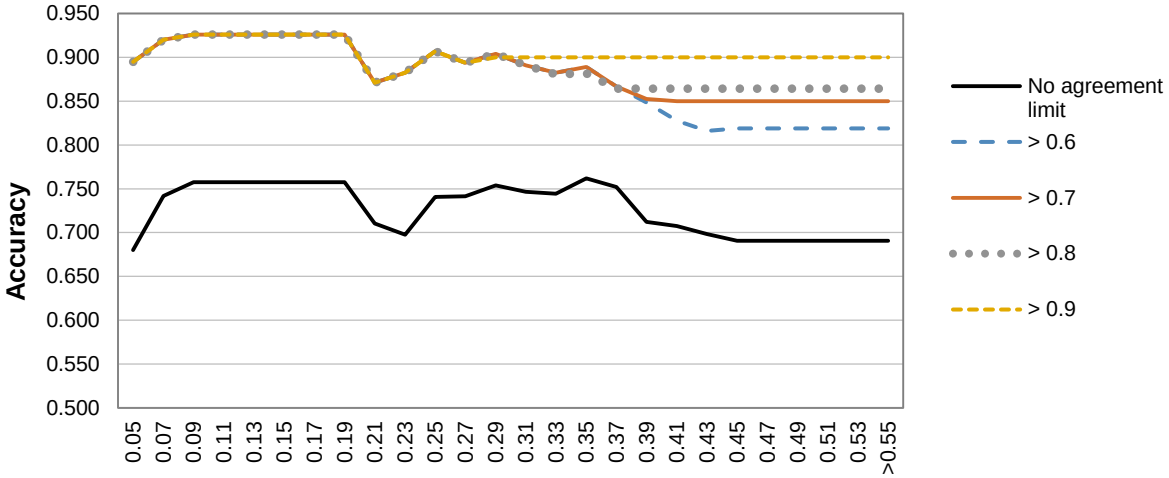


Figure 8.11. STD AD taking into account different agreement levels in the P-gp validation dataset.

To demonstrate the utility of RDN, consider one of the compounds with the lowest ensemble STD scores in our test set (Pemirolast, shown in Fig. 12, has an STD of 0.0284). According

to its STD score, this compound would be deemed very reliably predicted, however it is actually systematically mispredicted. In contrast to STD, the RDN applicability domain only covers this compound at around 70% data coverage. As a result RDN is effectively able to overcome this systematic bias and correctly identify this as a lower-reliability prediction.

As RDN AD describes a consistent relationship between distance-to-model (or RDN distance) and accuracy in two external datasets, it should be used as a measure of prediction confidence across the chemical space, rather than merely a single point AD threshold where some compounds are included while others are excluded. Hence, instead of assigning compounds as in- or out-of-domain, they should be associated with different prediction confidences. This is a more sensible use for the AD, as it would be up to the end user to select the maximum acceptable error rate level.

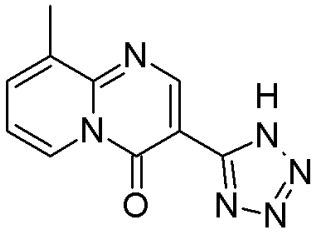


Figure 8.12. Example of an external set compound (Pemirolast) whose prediction is misleadingly deemed reliable when using the STD method. However, the RDN correctly associated this with low-reliability prediction, which matches the misprediction outcome observed for this compound.

Furthermore, as shown by RDN and, to a lesser extent, by STD (Figures 8.8 and 8.10), this continuous AD characterization allows mapping the reliability landscape across the data. This can be used to identify problematic regions in the model, which is more productive than merely accepting or excluding predictions (as in the leverage AD, for example). For example, using Figure 8.8 B the predicted P-gp queries that fall in regions up to 13%, and between 22-27% of included data (which can be traced back to an underlying Euclidean distance threhsold) are expected to be more reliably predicted according to the AD profile. The AD profile also shows that from 70% inclusion onwards, there is a much higher probability of compounds being mispredicted.

Additionally, the impact of the minimum requirement for the amount of training neighbours was investigated (ranging between 2 and 30 as described in the Methods section 8.3) and the results revealed no benefit from increasing the amount of neighbours (see Appendix V, Supplement A5.1 “Impact of the minimum required number of training neighbours”).

8.4.4. Complementary Analysis with Other AD: Diagnosing Mispredictions

Descriptor range has been used as a simple way of defining the applicability domain of a QSAR model. Here, in order to identify whether mispredictions are more commonly found outside the chemical span of the model, the descriptor range of the training set compounds at each of the branches in the decision tree model was computed. This strategy was previously proposed by Tong et al. (Tong et al., 2004), however, the descriptor range was limited to the instances actually passing through each of the tree branches, instead of considering the descriptor range of the entire dataset. The rationale behind this approach is that a given tree ramification may, for example, establish that class 1 has MW > 100 $g.mol^{-1}$ and class 2 has MW ≤ 100 $g.mol^{-1}$, which are one-sided limits. This means that a query with MW = 50 $g.mol^{-1}$ is able to pass through that node even though the training cases that pass through the same node have MW ranging [70-100]. In reality, this compound is outside the range "known" by the trained model, and will be detected as such by this approach (process illustrated in Figure 8.13). Curiously, the 62 test instances that fell outside the respective branch's descriptor range were associated with 72.6% accuracy, while the compounds inside the descriptor range showed 67.7% accuracy. This shows that falling outside training range is not necessarily the cause of misprediction. This justifies and further supports the use of an AD, like RDN, that identifies possible problematic regions within the data.

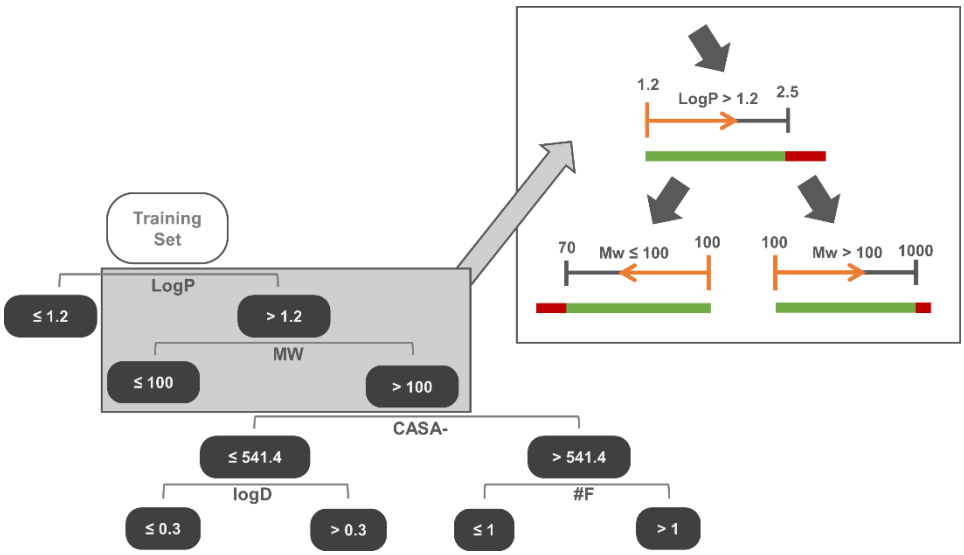


Figure 8.13. Schematics of the branch span assessment.

On the other hand, a method such as KDE, which is one of the most sophisticated AD approaches known for being able to detect empty regions in the data (Sahigara et al., 2012), also shows marked unpredictability in new data (Figure 8.14). Its utility is based on the

expectation that empty or less populated regions equate to weaker predictive performance due to insufficient chemical information. Figure 8.14 shows that the two external sets show different profiles (taking into account a comparison between the slopes of equivalent segments of both curves). This suggests that even looking at the inner space in descriptor range (which is the case with the KDE method), as opposed to looking at the descriptor range, does not appear to be sufficient by itself, as density appears to relate to predictive accuracy in a non-robust manner (Figure 8. 14). However, the figure still shows some level of correlation between density and predictive performance. Low percentage of data coverage indicates higher density thresholds in the density plot across the first principal component (used to calculate the density distribution model), and as this threshold is decreased (the AD boundaries get expanded) there is an overall trend of decreasing accuracy. Nevertheless this is still a very rough trend, and the fact that accuracy does not evolve in the same manner in both datasets, as data coverage is increased, indicates that addressing data density is not sufficient as a standalone AD measure, but it could be a useful parameter towards the characterization of a model's AD. This corroborates the inclusion of this property in the RDN algorithm.

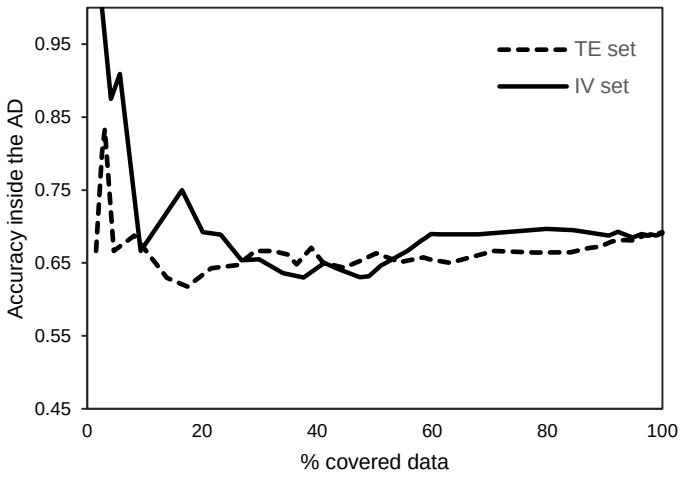


Figure 8.14. KDE results on validation and test sets of the P-gp dataset.

8.4.5. Evaluation of RDN on Benchmark Datasets

To validate the utility of RDN, this was applied to two previous models built from benchmark data, Ames and CYP450. Note that the two benchmark datasets were modelled using neural network training, while the P-gp data was modelled with a decision tree method. Additionally, recall that the same feature selection method was used for all AD methods across all datasets (ReliefF top 20 features).

Both benchmark modelled datasets resulted in a smooth, decreasing curve of accuracy vs percentage of included data in the AD with RDN (which directly translates into distance to the model) (Figures 8.15 and 8.16). Furthermore, the shape of the curve in the two external datasets within each benchmark dataset is similar. In addition to RDN, Figures 8.15 and 8.16 show that STD and dk-NN also generate curves of similar shape for the two external sets, however this was not the case for KDE. This reinforces the need to test a model's AD in two different sets of data.

The main difference between RDN and STD with respect to the Ames model was that RDN profiles differed only in absolute accuracy values and maintained a similar overall curve shape for the two external sets, whereas STD revealed a significant difference in shape between the two curves at the core of the AD. This is very likely due to systematic bias in the model, which produces agreeing predictions in the ensemble which are consistently incorrect (i.e. a low STD for incorrect predictions). As in the RDN method, both precision and bias are accounted for, this shortcoming has been overcome.

For CYP450 a similar overall performance to that with Ames has been obtained. Moreover, in this case, both external subsets showed very similar absolute accuracy values. STD performed also very reliably with CYP450 but, once again, there is more oscillation of accuracy near the core of the model than with RDN. This oscillation is however not so marked that it would lead one to question STD's robustness across other data. However, this is another example of a possible systematic bias that the ensemble STD could not overcome.

Results from both datasets confirm the validity of RDN as a method to appropriately define the applicability domain of a QSAR, by allowing a robust mapping of local predictive reliability across chemical space. Recall that this AD technique is completely independent from the model, and the AD is established solely using the training set. New predictions are merely sorted into different regions of the AD landscape after span of coverage around the training set has been set, at each iteration of the algorithm. The fact that correctly predicted instances show higher probability of being found near the training instances that are less biased and more precisely captured by the QSAR model demonstrates that, as theoretically expected, the reliability of a neighbourhood is inherited by its occupiers.

Furthermore, the independent role of density with respect to determining predictive reliability can be assessed by dk-NN and KDE as both sort the data according solely to density, where dk-NN does it at a local level, whereas KDE does it on a global scale. According to Figures 8.15 and 8.16, both KDE and dk-NN methods fail to achieve a descending level of accuracy with distance from the model's core. In addition, in both Ames and CYP450, the two different

external subsets show different profiles, indicating that density and predictive performance vary unpredictably with respect to each other. As with the Pgp model, the Ames model also shows an overall slight descending trend with KDE and dk-NN. This supports the hypothesis that utilizing density information (both local and global) could play a role in the determination of a robust AD.

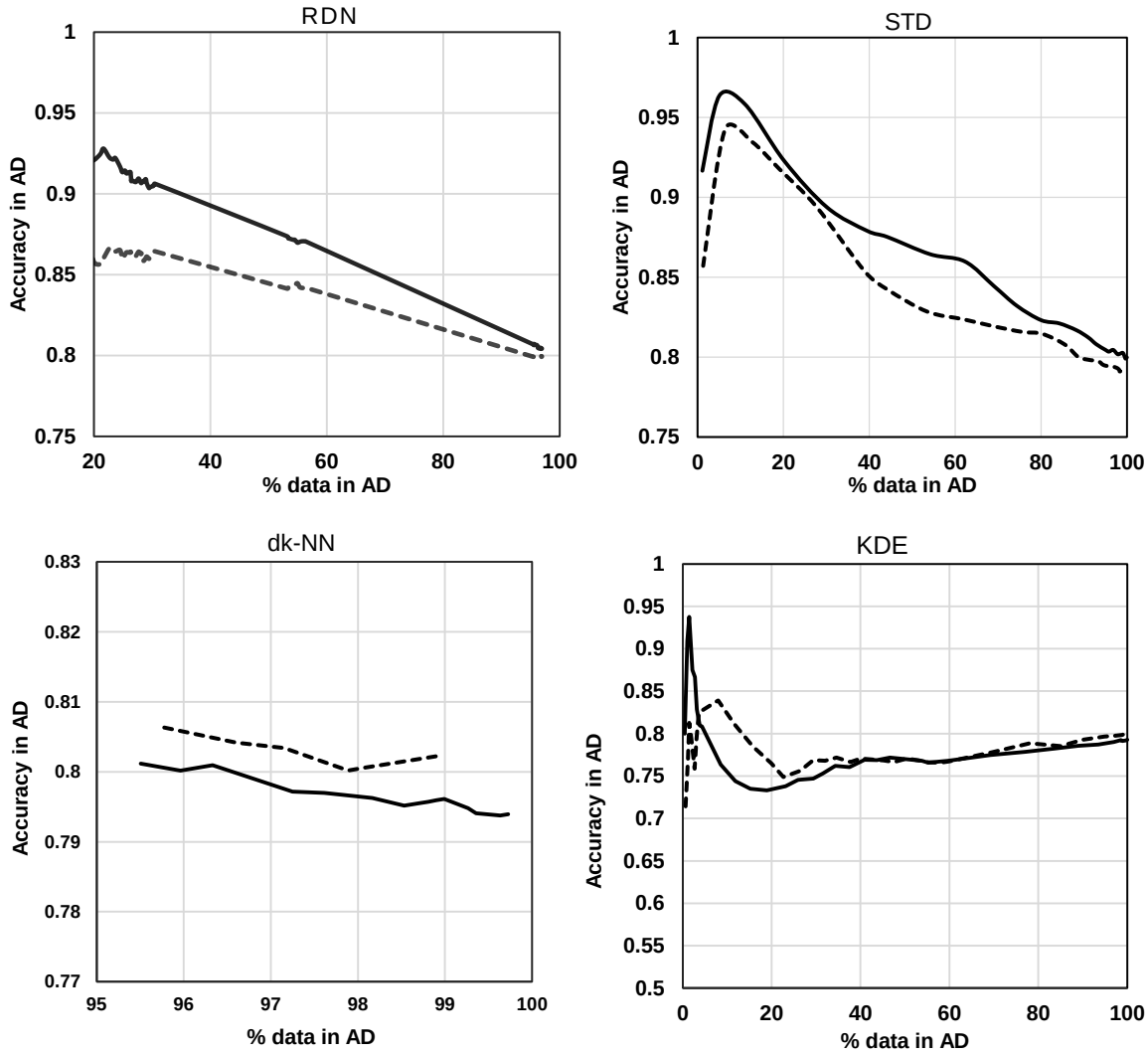


Figure 8.15. All four AD methods applied to the Ames model. Each of both lines in each graph corresponds to the same partition of the test set. Each line type represents one of the two external test sets from the Ames dataset.

On the other hand, the fact that the two CYP450 external datasets show quite different profiles with KDE, and this same technique has very different outcomes between all three datasets, indicates that this method is not reliable as a standalone measure for AD determination and there may be other factors that should be taken into account. While global density appears to have an unpredictable role in predictive reliability, one cannot conclude that density has no role in the establishment of an AD, as when it is addressed at

a local level in the dk-NN method, it shows very low resolution at the core, which might be hiding meaningful correlations with accuracy.

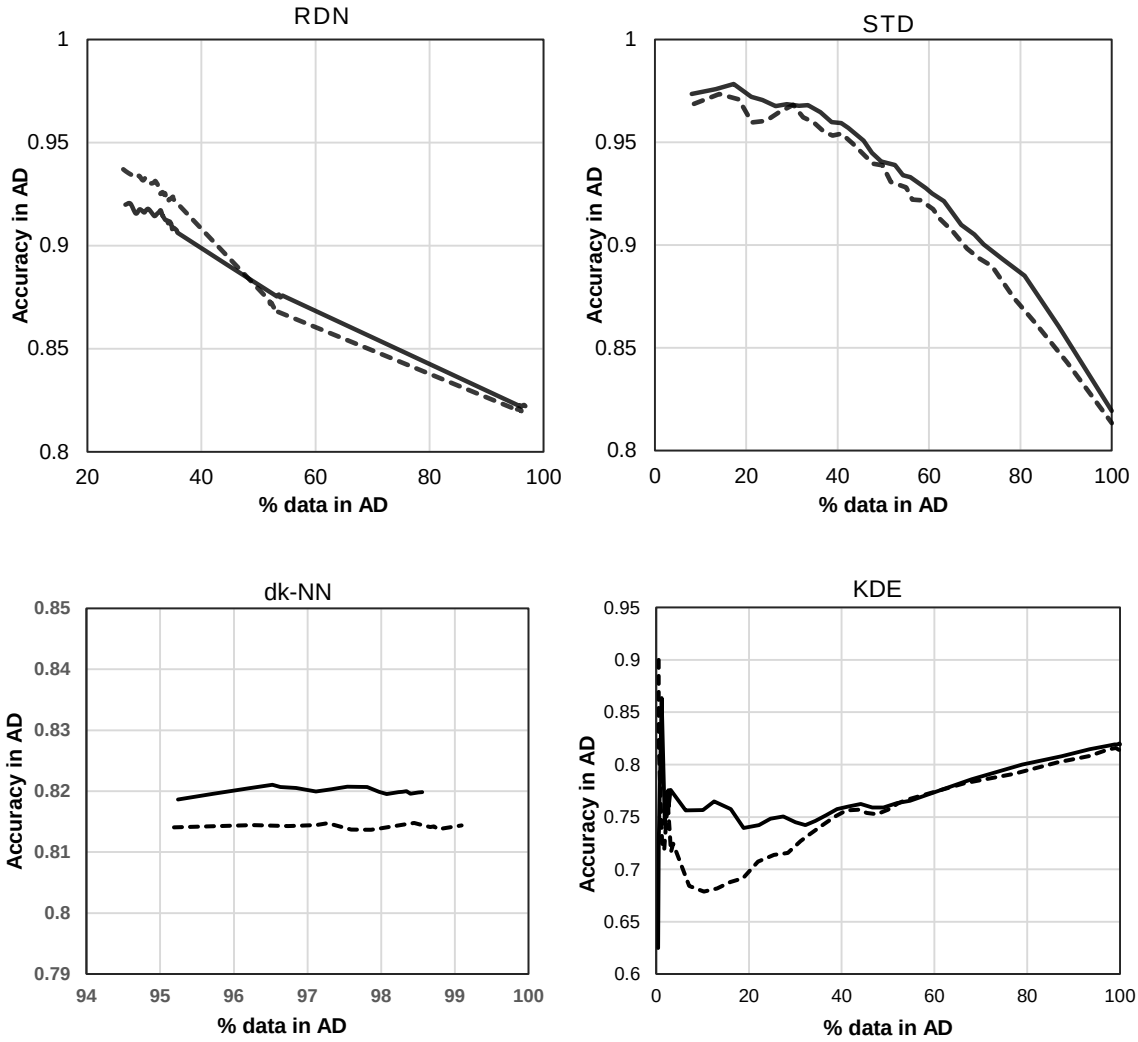


Figure 8.16. All four AD methods applied to the CYP450 model. Each of both lines in each graph corresponds to the same partition of the test set. Each line type represents one of the two external test sets from the Ames dataset.

8.4.6. Assessment of the AD Quality using a Scoring Function

Here a scoring function is proposed to numerically measure the efficiency of an AD curve (see the Methods section 8.3). Using this function leads to the same conclusions obtained from visual analysis of the AD profiles (scores are summarized in Table 8.1). According to the AD scoring function, Ames and CYP450 show more similar external set curves with RDN than with STD, which indicates that RDN is in general a more robust method for AD profiling. On the other hand, KDE obtained the worst (highest) score in all 3 models. Despite

what was previously established regarding the value of RDN, here the quality score points to the superiority of STD for the P-gp dataset. Recall that the quality score favours descending, smooth curves, and indeed STD has a smoother profile; however, RDN has the advantage of robustly locating poor quality regions (as discussed earlier). This shows that the scoring function may not necessarily follow the qualitative assessment of the AD profiles. Note that this study does not claim that RDN performs better than STD in all possible scenarios and datasets; instead, as with model development, the best AD method must be evaluated and the best method adopted in a case-by-case situation within every modelling effort. It is possible that some datasets suffer more from the effects of bias and hence they would benefit from RDN to overcome the systematic bias aspect of the STD method. This could explain why Ames and CYP450 models showed a very strong correlation between accuracy and distance to training space using RDN, and the P-gp model shows a poorer trend.

As explained in the Methods section 8.3, in the calculation of the scoring function, the impact of any given sub-segment of the AD curves is corrected for the amount of data it is associated with. Consequently, even though visually all points in an AD curve carry the same weight, the proposed scoring scheme allows assigning the correct weight to each point according to the number of implicated instances. As a result, even though, in a comparison between CYP450-STD and CYP450-RDN, the AD characterization of the models with STD appears to be as robust as the RDN in the AD profile figures, STD is in fact associated with more data being located in uncertain regions of chemical space.

Table 8.1. Summary of AD score across all three models studied. Lower AD scores indicate a better scenario, translating into higher similarity to an ideal AD curve (smooth and decreasing trend of accuracy as a function of the AD span), and it also translates into a closely matching pair of two external set curves (which translates into a higher level of robustness). The lowest scores for each dataset are highlighted in boldface.

	AD Score			
	RDN	STD	dk-NN	KDE
P-gp	4.40	**2.79**	6.82	8.14
Ames	**1.29**	1.92	4.48	9.26
CYP450	**1.01**	2.85	7.84	13.00

In order to support the validity of this AD robustness score, it is worth analysing the contribution of simpler measures (or concepts) that are incorporated in the newly proposed score. The details of such analysis are available in Appendix V (Supplement A5.2

"Complementary assessment of simpler curve similarity measures"), where it can be seen that none of the two parameters that constitute the proposed score, i.e., the pairwise similarity and the absolute difference between the curves, are sufficient on their own in assessing the quality of an AD profile; and the proposed scoring function is the most appropriate measure of AD robustness.

The fact that P-gp data is smaller and very noisy makes it more difficult for AD development. The P-gp data generated a poorer model (inferior test accuracy) (Aniceto et al., 2016b), with a higher rate of mispredictions than Ames and CYP450 models, which makes the task of defining a smooth AD profile considerably harder. The noise in the P-gp data comes from the variable threshold used in various sources to consider a compound as being a substrate (Broccatelli, 2012), as well as the very large level of experimental uncertainty (Bentz et al., 2013). Furthermore, P-gp binding is notably known as being a very complex phenomenon driven by outstanding polyspecificity (Chufan et al., 2015), which makes it naturally prone to error or bias in the experimental data.

8.5. Conclusions

The utility of a QSAR relies on the theoretical assumption of a smooth relationship between independent features and the dependent variable (Maggiora, 2006), which allows its use for interpolations. However, as in reality the model's landscape is not entirely smooth, it is crucial to map rugged regions across chemical space, since identifying these regions is the only way of assuring that the model can be safely used for future predictions (Krein et al., 2012). The applicability domain establishes where the Structure-Activity relationship is smooth (i.e., where the dependency between structure and property holds). These rough "patches" in the structure-activity landscape could be due to input errors, abrupt changes in activity/property known as activity cliffs, or lack of chemical coverage due to data scarcity. It is proposed here that the adequate feature set optimized for the characterization of the AD can, in theory, reveal the problematic regions if the AD is optimized using external sets. By testing the AD performance with new data (external set), this increases the probability of having compounds falling in such "unseen" regions of structure-activity. As a result, the poor ability to predict these compounds will pinpoint the locations where the model should not be used. To address this issue, this study introduced a novel AD characterization method that considers the impacts of local data density, as well as the precision and robustness of predictions across the chemical space. In addition, the role of feature

selection paired with the AD technique was also addressed, challenging the usual inheritance of features previously selected for the model development.

The new AD technique proposed in this work, named Reliability-Density Neighbourhood (RDN), is a hybrid technique, joining features from a density k-NN approach (which is here referred to as dk-NN) and the standard deviation of an ensemble model (named STD), as well as additional novel features like bias correction. The RDN AD allows taking into account: (1) sparse regions by mapping data density, as well as (2) local precision and bias. At the same time, this method was paired with ReliefF, which selects a set of molecular descriptors optimized to allow maximum separation between the classes to be predicted by the model. This method was applied to three different QSAR datasets and was compared with other established AD methods. Using the RDN AD allowed to improve the original distance-to-model method (dk-NN), which can be regarded as a simpler version of RDN. This improvement was visible through the increase of the accuracy at the core of the AD. RDN showed to be a robust AD technique that maintains an expected profile where performance degrades with increasing distance to the model in an external set. This technique showed overall better performance in comparison with the established STD method, as well as when compared with KDE, across all three datasets with a very strong correlation with accuracy.

The results presented here indicate that a given applicability domain needs to be assessed by the use of more than one external dataset to investigate the robustness of the AD. The two external sets can be compared in terms of accuracy vs distance-to-model profiles to indicate the reliability of a proposed AD. In this chapter, a scoring function to assess the quality of a given AD was also presented. The scoring function takes into account both robustness and the strength of the correlation with accuracy. As a result the assessment of robustness is proposed as standard procedure during the characterization of an AD, which can be done by evaluating the similarity of the relationship between accuracy and an AD measure for the two external subsets. This is a paramount aspect to take into account; without this there is no indication that a given AD can maintain its established accuracy profile across chemical space with new data.

This work challenges the common notion that either the QSAR model's features or the entire feature set must be utilized for the establishment of the AD, and proposes that a separate feature selection task should be performed specifically for AD development. Due to its particular characteristics, ReliefF has been proposed as a very effective algorithm for this. The results of this work showed that the feature set leading to the highest predictive performance is not necessarily the most adequate feature set for AD characterization. The

proposed implementation of a feature selection routine using ReliefF showed to be successful in mapping accuracy across the structure-activity landscape.

Overall the RDN technique was shown to effectively map prediction reliability across a QSAR model's chemical space, and shown to be a useful tool to guide users on their decision regarding compound prioritization, thus promoting the user's trust with the utility of the QSAR model itself. This work helps reinforce the central role of AD characterization in any modelling workflow, as it demonstrates the importance of a thorough implementation and characterization of the AD.

9. Conclusions and Future Perspectives

This final chapter gives an overview of the research presented in this thesis, as well as its relevance and novelty to the field of chemoinformatics.

As discussed in the Introduction (Chapter 1), the ability to properly screen compounds according to their ADME amenability has been, and still is, a key factor in reducing late stage drug attrition and making the drug development process more economically viable. Computational models have enabled the improvement of the capability for ADME screening, among which QSAR models show a very attractive balance between allowing very high throughput and still producing relatively good predictive performance. QSAR modelling additionally offers the possibility of deconvolution of desired or undesired chemical patterns/ scaffolds (with respect to PK) from the measured endpoint, whose information can be fed back into the drug development process to help guide current or future campaigns.

Among the different ADME properties that determine drug failure and contribute to the attrition rate, this thesis has focused on Volume of Distribution (Vd), and the prediction of this variable has been extensively explored in the past decade. It is well established that this variable depends on an array of both physicochemical and physiological factors however, while physiological models (PBPK models) have been trying to harness both sources of information, the statistical (machine learning) modelling of this property has been mostly explored using chemical features as the only source of input. Besides the work by Freitas et al. (Freitas et al., 2015) that uses tissue partition data as input (alongside molecular descriptors) in a machine learning QSAR model, QSAR models in the literature have tipically used phospholipid binding and plasma protein binding as physiological input. Alterantively, Vss has been indirectly predicted from physiological information by Paixão et al (Paixão et al., 2014), where physiological descriptors are used to model tissue partition across the main tissue contributor of Vss. Tissue contributions are have, in turn, been summed to calculate Vss. As a result, as of writing this thesis there was still no data-driven (QSAR) approaches to modelling Vss that attemped to use more, and more complex information of physiological information as input. As a contribution to filling this gap, a more sophisticated approach to better capture the distribution process was designed.

Borrowing from empirical knowledge and the physiological models published in the last decade, the overall hypothesis in this research is that Vd (measured as Vss) can be better modelled using a statistical approach accompanied by the incorporation of input physiological descriptors alongside physicochemical features. This hypothesis is derived from the knowledge that several different physiological mechanisms directly affect

distribution - in fact, formally Vss is the net effect of the conjugation of such driving mechanisms. This approach is further encouraged by the positive impact from directly accounting for such processes, reported for instance by Freitas et al. (Freitas et al., 2015).

To test this hypothesis, transport data from two main families of transporters – the ABCs and the SLCs - was selected to be used as input features. However, as there is small-to-no overlap between the publicly available data from these transporters and the benchmark dataset for Vss in humans (published by Obach et al (Obach et al., 2008), which exaplins the lack of studies using such kind of information as input to model Vss. This was overcome by imputing missing transport data using predictions produced by previously trained QSAR models on various key transporters. Using such imputation approach shares the same principle behind multi-label modelling, particularly the classifier chain approach, where, if two properties (or labels) share any form of correlation, it might be benefitial to model one property and use the resulting predictions as a features of the second property. Similarly, in the case of this work, predictions for transporter data were taken as a feature in the modelling of Vss. To do so, a previous step of ABC and SLC transport modelling was carried out.

The ATP-Binding Cassette (ABC) transporters were modelled first and this effort was discussed in Chapter 4. To do so, substrates and non-substrates from four of the ABC transporters (BCRP, P-gp, MRP1 and MRP2) with the highest clinical significance were gathered. Given the overlap observed between transporters and the existence of instances where these transporters show cooperation or redundancy in similar partition processes in certain tissues, a multi-label approach called Classifier Chains was employed to ascertain whether there is any correlation between these four different efflux processes, and if so, to use it in order to aid the learning of their respective structure-activity relationships. In this approach, decision trees for predicting substrates of the transporters were trained iteratively, and the learned information was passed forward, and used as a predictor to train models for the following transporters, in a chained process. The CC model showed evidence of learned information from certain transporters being effectively picked into the decision trees, and some improvement of predictive performance was observed, when compared to a scenario where transporters where modelled independently. The work reported in that chapter demonstrated evidence of the hypothesized correlation between the four ABC transporters, as well the evidence for the feasibility of using multi-label classification to harness such correlations.

With this work, a novel multi-label classification approach to model ABC efflux data was porposed, being only employed in another two works (Ose et al., 2016, Montanari et al., 2016) that were published during the same time-frame as this work. However, Ose et al is not an explicit multi-label work and the authors also do not explore label interaction (which is arguably the main gain from using multi-label classification); as for Montanari et al, they addressed inhibitory data instead, and only two transporters were modelled. Exploring the multi-label classification fo the ABCs opened new avenues worth investigating. One of the most obvious aspects to explore further (which was also one of the main limitations of this study) would be addressing and optimizing label order in the classifier chain model, as this aspect has a high impact over the predictive power yilded from a CC model. Even though there are alternatives available which allow optimization of label order, in this case an exhaustive exploration would be feasible as 4 labels (or transporters) can be rearranged in a total of 24 permutations. It would be interesting to analyse how the optimal label ordering obtained from a purely data-driven approach compares to the physiological relationship between transporters. Conceivably, labels with higher ligand promiscuity or diversity may benefit from receiving information from more specific labels as, in the first case, structural information alone might be less able to differentiate substrates from non-substrates, owed to polyspecificity of binding. Another aspect worth investigating would be the use of physiological informatin in the modelling of the different ABC transporters. As explored in Chapter 7, physiological descriptors show different profiles of tissue expression, and exploring ways of harnessing this information towards (1) adjusting transporters' contributions and order during the optimization of a classifier chain model order, as well as (2) improving each label's prediction performance. In addition to tissue expression, the relationship between phospholipidosis and ABC transporter could yield very useful information in toxicity prediction as, on one hand, phospholipidosis occurs mainly in certain tissues like the liver, lung, brain and kidney and, on the other hand, different populations of transporters differ between tissues. Additionally a future follow-up to the current work would also involve testing the ABC multi-label model in larger datasets, as a way to better establish their actual predictive value, as well as attempting to use more powerful machine learning algorithms such as random forest or support vector machine.

Next in the workflow of this thesis the SLCs were modelled, in a similar approach of that used for the ABCs, with the findings being presented in Chapter 5. Substrates and non-substrates for 6 SLC members – OATP1B1, OATP1B3, OATP2B1, OATP1A2, PEPT1 and OCT1 – were modelled using multi-label classification, specifically a classifier chain model. However here, inspired from the shortcoming of the ABC multi-label chapter where label order was established on an empirical basis, the order in which the transporters are

modelled in the classifier chain was optimized in the SLC model, through exhaustive exploration of all combinations, permutations and chain lengths of the six transporters. Also, now the machine learning algorithms were also optimized for each transporter label, which was another shortcoming in the ABC model. In this work, all tested classifiers were tree-based algorithms (random forest, boosted trees or decision trees) which were optimized for each transporter. This effort yielded much stronger evidence of transporter interaction than the ABC project, and the best CC model was found to be one containing all six transporters, showing several accounts of learned information from a previously trained transporter model being used as a predictor.

One of the main aspects of novelty contributed through this work was the fact that it proposed several links between different pairs of transporters, some of which are not obvious and have not yet been reported. Additionally, this is the first attempt to apply multi-label classification to SLC data, which contributes to disseminating this technique among the chemoinformatics comunity as a viable QSAR technique to explore and uncover new potentially related endpoints. Considering the new links proposed among the SLCs, one interesting follow-up to this work would be to challenge this with extended external data, especially considering that some of the datasets are rather small for one to be able to draw robust conclusions. Finaly, as suggested for ABCs, it would be relevant to explore the potential benefits from addressing tissue expression in optimizing label order in a multi-label model. In this work, label order was optimized using and internal validation set (i.e. label order is selected according to the highest prediction performance achieved in the internal validations et) however, as generally known and as systematically explored in a recent study by Martin et al across a wide range of bioactivity dataset, relying upon a random subsample of the dataset to measure performance might produce a skewed assessment of the models performance, when compared to a realistic external dataset. As a result, exhaustively testing all label arrangements might produce a best label ordering that is distorted by biases in the dataset; such biases might be controlled by using other sources of information such as tissue localization profiles of different proteins, as this data offers a different perspective of transporter relationships and might provide some implicit information on hierarchical links between transporters.

For both the ABC and the SLC works, it would also be interesting to explore criteria for data quality and how this affects the multi-label performance, as data from transport assays result from an arbitrary threshold that separates substrates and non-substrates. In addition to this, incorporating an applicability domain filter into both multi-label models would be beneficial, and a novel concept to explore within the multi-label modelling field. This is especially

important in classifier chains as it might help control the main limitation in this method - the propagation of errors produced in each single label.

The work in Chapter 6 drew from work carried out in both Chapter 4 and 5, as in this chapter the central hypothesis of this thesis was tested by exploring the feasibility of using physiological descriptors in the modelling of Vss, alongside chemical features. These physiological descriptors included ABC experimental data completed with predicted output by the CC model in Chapter 4, as well as SLC experimental data completed with prediction from the CC model in Chapter 5. In addition, drug-induced phospholipidosis data was modelled and experimental data was completed with predictions, as with the ABC and SLC predictors. Predictions were used in the form of output probabilities, to allow the machine learning algorithm to distinguish between predictions of higher and lower quality. After an optimization of parameters used in the modelling, where different candidate models were built using different feature selection methods, different machine learning algorithms and different feature types, the best model contained physiological descriptors that included examples of all three types of physiological sources of data: phospholipidosis, ABC and SLC features. After some benchmark comparison with other works that only used chemical predictors, the best model in this work showed to yield superior performance in fixed external datasets. Additionally, using the feature set of the best model re-applied on a more recent, larger Vss dataset also showed that using these physiological descriptors consistently improves predictive performance. Such observations represent an important improvement of the ability to screen and locate compounds in an early stage which are more prone to have extreme (hence problematic) expected Vss values. In addition, and as a bi-product of producing both Vss predictions and efflux/uptake predictions simulataneously for ever compound, this modelling scheme allows a more detailed profiling of compounds that are selected, as one might prefer a compound that is in the same range of Vss values as 10 other compounds, but does not show to undergo efflux and is only expected to undergo uptake. This type of profiling is made easy due to the integration of various sources of information in the model.

This work's main novel contribution consisted of exploring the value of 1) using a variety of physiological processes as input features of Vss and 2) using predicted data to enrich the physiological variables improve the modelling of Vss. The second point is of especial importance as it allows using physiological information to inform Vss in a high throughput setting, thus overcoming the limitation imposed by availability of experimental data such as efflux/uptake and drug-induced phospholipidosis. Conceivably, this approach can be extended to other physiological properties of interest, which may potentially aid greatly the ability to predict Vss. Additional future work that could stem from the work presented here

includes carrying out the full battery of tests applied during model optimization (which correspond to all tested variations of the modelling conditions mentioned earlier) but applied to the larger, recently published Vss dataset (Lombardo and Jing, 2016). Additionally, testing different probability cutoffs for the inclusion of the predicted data would be important to establish how this affects the model performance, and whether some data should be excluded (e.g., probabilities between 0.4 and 0.6, where 0 determines maximum confidence in being a non-substrate and 1 determines maximum confidence in being a substrates) to reduce noise and improve the classification task. In other words, and in the same line of what was suggested for the ABC and SLC models, it would be beneficial to explore the applicability domain of each predicted variable, and incorporate a reliability/confidence filter to every physiological features before it is used as a predictor in the training of a Vss model.

Chapter 7 consisted of a further exploration of the premise in chapter 6, where Vss was also modelled with ABC, SLC and phospholipidosis data, but the transporter responses used in the modelling in the previous chapter underwent correction using tissue expression data in a large variety of tissues. As a result, a 0-1 scale attributed to each of the transporter variables was converted into the net sum of transport extent in a collection of tissues. This was prompted by the notion that being a substrate with 90% probability for a transporter that exists in 10% of tissue mass has a different impact in distribution from being a substrate with 90% probability for a transporter that is expressed in 80% of tissue mass. This expression correction revealed to yield a small but consistent improvement in predictive performance from the previous best model (in Chapter 6), which was already and improvement from benchmark conditions. As seen in the previous chapter, several physiological descriptors were spontaneously selected even when submitted to feature selection alongside all available molecular descriptors. In this instance, it might be interesting, as future work, to draw rules from physiologically-based models to create a more refined correction system. One key detail covered with both chapter 6 and 7 is the importance of the structure of the feature selection procedure. Usually reported research carries out one feature selection step typically performed on all descriptors in bulk however, this thesis shows that very different outcomes can result from running separate feature selection steps on different types of features.

Attention should be paid to the fact that beyond the fact that predictive performance has improved from all other QSAR works carried to date on Vss modelling, the performance values reported are actually very conservative. This is due to the fact that, contrarily to any other work on volume of distribution reported in the past, the current work is able to sort predictions according to their confidence is a very reliable manner. So, in practice, upon

employing a confidence threshold to predictions, a much lower error can be obtained for “confidence-passing” compounds.

Up until this point all experimental chapters have a common denominator concerning future proposed work which is tied to the characterization of each model’s applicability domain, particularly for the purpose of controling the quality of input features containing predicted data. Additionally, there was a concern throughout this thesis to define the confidence associated with new predictions output from all the different QSAR models. Given the central role of applicability domain characterization in validation for produced model and considering it is now deemed by the community as being as important as demonstrating high actual predictive performance, Chapter 8 reports the finding on the development of a novel applicability domain method named reliability-density neighbourhood, or RDN. This method showed the ability to sort external data according to the accuracy with which they are predicted, doing so by taking into account the local predictive bias and precision, as well as local density, across different neighbourhoods of the training space. Testing this method in two benchmark datasets showed excellent correlation between accuracy and span of coverage. The main contribution provided through this work is that it is the first applicability domain method that attempts to address bias in the predictions for unseen data. Other very successful methods to define applicability domain exist such as prediction standard deviation or conformal prediction, however these only addess precision.

Other novel contributions brought by Chapter 8 consist of the proposal of a new scoring function that quantifies the relative quality of the produced applicability domain profiles, and the systematic study of the impact of feature selection in applicability domain characterization. One of the possibilities for future work is testing different methods to establish local density, as currently this relies on Euclidean distances to nearest neighbours. One promising alternative to this could be computing distance using Tanimoto coefficient calculated from circular fingerprints. Additionally, as the Euclidean distance was calculated based on physicochemical features, other features such as fingerprints or fragments could also be explored alone and alongise.

Based on the research carried out in this thesis, it can be concluded that QSAR modelling, particularly using the multi-label learning approach, can aid in understanding and better predicting drug distribution processes, and it can provide new insight by addressing different, related endpoints as a whole. Multi-lable also showed to be useful in discovering new possible interactions of therapeutic targets, which can be used to generate new therapeutic options as well as generating new hypotheses for physiological pathways. Integrating different sources of information of both physiological and physicochemical nature has shown to contribute towards capturing volume of distribution and using both

types of information produces a mean prediction error inferior to when either chemical or physiological features are used alone. This approach can potentially be used to mine new relationships that may be of potential clinical relevance to the pharmaceutical industry. This work, while developed on relatively small datasets compared to the data size used in real world PK profiling in drug discovery, is a proof of concept for the usefulness of QSAR models trained with information from underlying processes that drive distribution, as a tool to improve the prediction of volume of distribution.

10. References

AGGARWAL, C. C. 2014. An Introduction to Data Classification. Data Classification, Hall/CRC.

ALEXANDER, D. L. J., TROPSHA, A. & WINKLER, D. A. 2015. Beware of R2: Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models. Journal of Chemical Information and Modeling, 55**,** 1316-1322.

ALFAROUK, K. O., STOCK, C.-M., TAYLOR, S., WALSH, M., MUDDATHIR, A. K., VERDUZCO, D., BASHIR, A. H. H., MOHAMMED, O. Y., ELHASSAN, G. O., HARGUINDEY, S., RESHKIN, S. J., IBRAHIM, M. E. & RAUCH, C. 2015. Resistance to cancer chemotherapy: failure in drug response from ADME to P-gp. Cancer Cell International, 15**,** 71.

ALLEN, T. J., MURPHY, A. J. & JANDELEIT-DAHM, K. A. 2015. RAGE Against the ABCs. Diabetes, 64**,** 3981-3983.

ANDERSON, N. & BORLAK, J. 2006. Drug-induced phospholipidosis. FEBS Letters, 580**,** 5533-5540.

ANICETO, N., FREITAS, A. A., BENDER, A. & GHAFOURIAN, T. 2016a. A novel applicability domain technique for mapping predictive reliability across the chemical space of a QSAR: reliability-density neighbourhood. Journal of Cheminformatics, 8**,** 69.

ANICETO, N., FREITAS, A. A., BENDER, A. & GHAFOURIAN, T. 2016b. Simultaneous prediction of four ATP-binding cassette transporters substrates using multi-label QSAR. Molecular Informatics, 35**,** 514–528.

ARTURSSON, P., MATSSON, P. & KARLGREN, M. 2013. In Vitro Characterization of Interactions with Drug Transporting Proteins. In: SUGIYAMA, Y. & STEFFANSEN, B. (eds.) Transporters in Drug Development: Discovery, Optimization, Clinical Study and Regulation. New York, NY: Springer New York.

BALLARD, P., BRASSIL, P., BUI, K. H., DOLGOS, H., PETERSSON, C., TUNEK, A. & WEBBORN, P. J. H. 2012. The right compound in the right assay at the right time: an integrated discovery DMPK strategy. Drug Metabolism Reviews, 44**,** 224-252.

BARBOUR, A., SCHMIDT, S., MA, B., SCHIEFELBEIN, L., RAND, K. H., BURKHARDT, O. & DERENDORF, H. 2009. Clinical Pharmacokinetics and Pharmacodynamics of Tigecycline. Clinical Pharmacokinetics, 48**,** 575-584.

BAUCH, C., BEVAN, S., WOODHOUSE, H., DILWORTH, C. & WALKER, P. 2015. Predicting in vivo phospholipidosis-inducing potential of drugs by a combined high content screening and in silico modelling approach. Toxicology in Vitro, 29**,** 621-630.

BENTZ, J., O'CONNOR, M. P., BEDNARCZYK, D., COLEMAN, J., LEE, C., PALM, J., PAK, Y. A., PERLOFF, E. S., REYNER, E., BALIMANE, P., BRÄNNSTRÖM, M., CHU, X., FUNK, C., GUO, A., HANNA, I., HERÉDI-SZABÓ, K., HILLGREN, K., LI, L., HOLLNACK-PUSCH, E., JAMEI, M., LIN, X., MASON, A. K., NEUHOFF, S., PATEL, A., PODILA, L., PLISE, E., RAJARAMAN, G., SALPHATI, L., SANDS, E., TAUB, M. E., TAUR, J.-S., WEITZ, D., WORTELBOER, H. M., XIA, C. Q., XIAO, G., YABUT, J., YAMAGATA, T., ZHANG, L. & ELLENS, H. 2013. Variability in P-Glycoprotein Inhibitory Potency (IC50) Using Various in Vitro Experimental Systems: Implications for Universal Digoxin Drug-Drug Interaction Risk Assessment Decision Criteria. Drug Metabolism and Disposition, 41**,** 1347-1366.

BERELLINI, G., SPRINGER, C., WATERS, N. J. & LOMBARDO, F. 2009. In Silico Prediction of Volume of Distribution in Human Using Linear and Nonlinear Models on a 669 Compound Data Set. Journal of Medicinal Chemistry, 52**,** 4488-4495.

BIAU, G. 2012. Analysis of a random forests model. Journal of Machine Learning Research, 13**,** 1063-1095.

BLUM, A. L. & LANGLEY, P. 1997. Selection of relevant features and examples in machine learning. Artificial Intelligence, 97**,** 245-271.

BOLÓN-CANEDO, V., SÁNCHEZ-MAROÑO, N. & ALONSO-BETANZOS, A. 2013. A review of feature selection methods on synthetic data. Knowledge and Information Systems, 34**,** 483-519.

BOLÓN-CANEDO, V., SÁNCHEZ-MAROÑO, N. & ALONSO-BETANZOS, A. 2015. A Distributed Feature Selection Approach Based on a Complexity Measure. In: ROJAS, I., JOYA, G. & CATALA, A. (eds.) Advances in Computational Intelligence. Springer International Publishing.

BOULESTEIX, A.-L., JANITZA, S., KRUPPA, J. & KÖNIG, I. R. 2012. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2**,** 493-507.

BOYER, S., BREALEY, C. & DAVIS, A. M. 2015. Attrition in Drug Discovery and Development. Attrition in the Pharmaceutical Industry. John Wiley & Sons, Inc.

BRANDSCH, M. 2013. Drug transport via the intestinal peptide transporter PepT1. Current Opinion in Pharmacology, 13**,** 881-887.

BREIMAN, L. 2001. Random Forests. Machine Learning, 45**,** 5-32.

BROCCATELLI, F. 2012. QSAR Models for P-Glycoprotein Transport Based on a Highly Consistent Data Set. Journal of Chemical Information and Modeling, 52**,** 2462–2470.

CARRIO, P., PINTO, M., ECKER, G., SANZ, F. & PASTOR, M. 2014. Applicability Domain ANalysis (ADAN): a robust method for assessing the reliability of drug property predictions. Journal of Chemical Information and Modeling, 54**,** 1500-11.

CARVALHO, A. C. P. L. F. & FREITAS, A. A. 2009. A Tutorial on Multi-label Classification Techniques. In: ABRAHAM, A. (ed.) Foundations of Computational Intelligence. Springer-Verlag Berlin Heidelberg.

CASCORBI, I., FLÜH, C., REMMLER, C., HAENISCH, S., FALTRACO, F., GRUMBT, M., PETERS, M., BRENN, A., THAL, D. R., WARZOK, R. W. & VOGELGESANG, S. 2013. Association of ATP-binding cassette transporter variants with the risk of Alzheimer's disease. Pharmacogenomics, 14**,** 485-494.

CASEY, J. R., GRINSTEIN, S. & ORLOWSKI, J. 2010. Sensors and regulators of intracellular pH. Nature Reviews Molecular Cell Biology, 11**,** 50-61.

CESAR-RAZQUIN, A., SNIJDER, B., FRAPPIER-BRINTON, T., ISSERLIN, R., GYIMESI, G., BAI, X., REITHMEIER, R. A., HEPWORTH, D., HEDIGER, M. A., EDWARDS, A. M. & SUPERTI-FURGA, G. 2015. A Call for Systematic Research on Solute Carriers. Cell, 162**,** 478-87.

CHAWLA, N. V. 2006. Many Are Better Than One: Improving Probabilistic Estimates from Decision Trees. In: QUIÑONERO-CANDELA, J., DAGAN, I., MAGNINI, B. & D'ALCHÉ-BUC, F. (eds.) Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment. Springer Berlin Heidelberg.

CHU, X., KORZEKWA, K., ELSBY, R., FENNER, K., GALETIN, A., LAI, Y., MATSSON, P., MOSS, A., NAGAR, S., ROSANIA, G. R., BAI, J. P. F., POLLI, J. W., SUGIYAMA, Y., BROUWER, K. L. R. & ON BEHALF OF THE INTERNATIONAL TRANSPORTER, C. 2013. Intracellular Drug Concentrations and Transporters: Measurement, Modeling, and Implications for the Liver. Clinical pharmacology and therapeutics, 94**,** 126-141.

CHUFAN, E. E., SIM, H.-M. & AMBUDKAR, S. V. 2015. Molecular Basis of the Polyspecificity of P-Glycoprotein (ABCB1): Recent Biochemical and Structural Studies. In: JOHN, D. S. & TOSHIHISA, I. (eds.) Advances in Cancer Research. Academic Press.

CIPOLLA, M. J. 2009. The Cerebral Circulation, Morgan and Claypool Publishers.

CONSONNI, V. & TODESCHINI, R. 2010. Molecular Descriptors. In: PUZYN, T., LESZCZYNSKI, J. & CRONIN, M. T. (eds.) Recent Advances in QSAR Studies: Methods and Applications. Dordrecht: Springer Netherlands.

COOK, D., BROWN, D., ALEXANDER, R., MARCH, R., MORGAN, P., SATTERTHWAITE, G. & PANGALOS, M. N. 2014. Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. Nat Rev Drug Discov, 13**,** 419-431.

CORDER, G. W. & FOREMAN, D. I. 2009. Comparing Two Related Samples: The Wilcoxon Signed Ranks Test. Nonparametric Statistics for Non-Statisticians. John Wiley & Sons, Inc.

CRIVORI, P., REINACH, B., PEZZETTA, D. & POGGESI, I. 2006. Computational Models for Identifying Potential P-Glycoprotein Substrates and Inhibitors. Molecular Pharmaceutics 3**,** 33-44.

CURRY, S. H. & WHELPTON, R. 2017. Introduction to Drug Disposition and Pharmacokinetics, John Wiley & Sons.

DANISHUDDIN & KHAN, A. 2016. Descriptors and their selection methods in QSAR analysis: paradigm for drug design. Drug Discovery Today, 21**,** 1291-302.

DANTZIG, A. H., HILLGREN, K. M. & DE ALWIS, D. P. 2004. Drug Transporters and Their Role in Tissue Distribution. Annual Reports in Medicinal Chemistry. Academic Press.

DE CERQUEIRA LIMA, P., GOLBRAIKH, A., OLOFF, S., XIAO, Y. & TROPSHA, A. 2006. Combinatorial QSAR Modeling of P-Glycoprotein Substrates. Journal of Chemical Information and Modeling, 46**,** 1245-1254.

DEHMER, M. & VARMUZA, K. (eds.) 2012. Statistical Modelling of Molecular Descriptors in QSAR/QSPR.

DEL AMO, E. M., GHEMTIO, L., XHAARD, H., YLIPERTTULA, M., URTTI, A. & KIDRON, H. 2013. Applying Linear and Non-Linear Methods for Parallel Prediction of Volume of Distribution and Fraction of Unbound Drug. PloS One, 8**,** e74758.

DEMIR-KAVUK, O., BENTZIEN, J., MUEGGE, I. & KNAPP, E.-W. 2011. DemQSAR: predicting human volume of distribution and clearance of drugs. Journal of Computer-Aided Molecular Design, 25**,** 1121-1133.

DESAI, P. V., SAWADA, G. A., WATSON, I. A. & RAUB, T. J. 2013. Integration of in silico and in vitro tools for scaffold optimization during drug discovery: predicting P-glycoprotein efflux. Molecular Pharmacology, 10**,** 1249-61.

DRAGOS, H., GILLES, M. & ALEXANDRE, V. 2009. Predicting the Predictability : A Unified Approach to the Applicability Domain Problem of QSAR Models. Journal of Chemical Information and Modeling, 49**,** 1762-1776.

EFRAT, A., FAN, Q. & VENKATASUBRAMANIAN, S. 2006. Curve Matching, Time Warping, and Light Fields: New Algorithms for Computing Similarity between Curves. Journal of Mathematical Imaging and Vision, 27**,** 203-216.

ELITH, J., LEATHWICK, J. R. & HASTIE, T. 2008. A working guide to boosted regression trees. Journal of Animal Ecology, 77**,** 802-813.

ERIKSSON, L., JAWORSKA, J., WORTH, A. P., CRONIN, M. T. D., MCDOWELL, R. M. & GRAMATICA, P. 2003. Methods for Reliability and Uncertainty Assessment and for Applicability Evaluations of Classification- and Regression-Based QSARs. Environmental Health Perspectives, 111**,** 1361-1375.

FAN, J. & DE LANNOY, I. A. M. 2014. Pharmacokinetics. Biochemical Pharmacology, 87**,** 93-120.

FAWCETT, T. 2006. An introduction to ROC analysis. Pattern Recognition Letters, 27**,** 861-874.

FDA 2012. Drug Interaction Studies - Study Design, Data Analysis, Implications for Dosing, and Labeling Recommendations. Draft Guidance.

FERREIRA, R. J., DOS SANTOS, D. J. V. A. & FERREIRA, M.-J. U. 2015. P-glycoprotein and membrane roles in multidrug resistance. Future Medicinal Chemistry, 7**,** 929-946.

FILIPPONE, E. J., CARSON, J. M., BECKFORD, R. A., JAFFE, B. C., NEWMAN, E., AWSARE, B. K., DORIA, C. & FARBER, J. L. 2011. Sirolimus-induced pneumonitis complicated by pentamidine-induced phospholipidosis in a renal transplant recipient: a case report. . Transplantation Proceedings, 43**,** 2792-2797.

FJODOROVA, N., NOVIČ, M., RONCAGLIONI, A. & BENFENATI, E. 2011. Evaluating the applicability domain in the case of classification predictive models for carcinogenicity based on the counter propagation artificial neural network. Journal of Computer-Aided Molecular Design, 25**,** 1147-58.

FLATEN, G. E., KOTTRA, G., STENSEN, W., ISAKSEN, G., KARSTAD, R., SVENDSEN, J. S., DANIEL, H. & SVENSON, J. 2011. In Vitro Characterization of Human Peptide Transporter hPEPT1 Interactions and Passive Permeation Studies of Short Cationic Antimicrobial Peptides. Journal of Medicinal Chemistry, 54**,** 2422-2432.

FOLEY, D. W., RAJAMANICKAM, J., BAILEY, P. D. & MEREDITH, D. 2010. Bioavailability through PepT1: the role of computer modelling in intelligent drug design. Current Computer-Aided Drug Design, 6**,** 68-78.

FORD, R. C., KAMIS, A. B., KERR, I. D. & CALLAGHAN, R. 2010. The ABC Transporters: Structural Insights into Drug Transport. Transporters as Drug Carriers. Wiley-VCH Verlag GmbH & Co. KGaA.

FOURCHES, D., MURATOV, E. & TROPSHA, A. 2010. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. Journal of Chemical Information and Modelling, 26**,** 1189-204.

FOWLER, PHILIP W., ORWICK-RYDMARK, M., RADESTOCK, S., SOLCAN, N., DIJKMAN, PATRICIA M., LYONS, JOSEPH A., KWOK, J., CAFFREY, M., WATTS, A., FORREST, LUCY R. & NEWSTEAD, S. 2015. Gating Topology of the Proton-Coupled Oligopeptide Symporters. Structure, 23**,** 290-301.

FRANKE, R. M., SCHERKENBACH, L. A. & SPARREBOOM, A. 2009. Pharmacogenetics of the organic anion transporting polypeptide 1A2. Pharmacogenomics, 10**,** 339-344.

FREITAS, A. A. 2013. Comprehensible classification models: a position paper. ACM SIGKDD Explorations Newsletter, 15.

FREITAS, A. A., LIMBU, K. & GHAFOURIAN, T. 2015. Predicting volume of distribution with decision tree-based regression methods using predicted tissue:plasma partition coefficients. Journal of Cheminformatics, 7**,** 6.

FUJITA, T. & WINKLER, D. A. 2016. Understanding the Roles of the "Two QSARs". Journal of Chemical Information and Modeling, 56**,** 269-274.

FUNK, R. S. & KRISE, J. P. 2013. Cationic amphiphilic drugs cause a marked expansion of apparent lysosomal volume: Implications for an intracellular distribution-based drug interaction. Molecular Pharmaceutics, 9**,** 1384–1395.

GALATHIYA, A. S., GANATRA, A. P. & BHENSDADIA, C. K. 2012. Improved Decision Tree Induction Algorithm with Feature Selection, Cross Validation, Model Complexity and Reduced Error Pruning. International Journal of Computer Science and Information Technologies, 3, 3427-3431.

GALAR, M., FERNANDEZ, A., BARRENECHEA, E., BUSTINCE, H. & HERRERA, F. 2012. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42**,** 463-484.

GANTA, S., SHARMA, P. & GARG, S. 2008. Permeability assessment. In: GAD, S. C. (ed.) Preclinical development handbook: ADME and Biopharmaceutical Properties. John Wiley and Sons.

GANTNER, M. E., EMILIANO, M., IANNI, D., RUIZ, M. E., TALEVI, A. & BRUNO-BLANCH, L. E. 2013. Development of Conformation Independent Computational Models for the Early Recognition of Breast Cancer Resistance Protein Substrates. BioMed Research International, 2013**,** 863592.

GEDECK, P., KRAMER, C. & ERTL, P. 2010. Computational analysis of structure-activity relationships. Progress in Medicinal Chemistry, 49**,** 113-60.

GHAFOURIAN, T., BARZEGAR-JALALI, M., DASTMALCHI, S., KHAVARI-KHORASANI, T., HAKIMIHA, N. & NOKHODCHI, A. 2006. QSPR models for the prediction of apparent volume of distribution. International Journal of Pharmaceutics, 319**,** 82-97.

GHAFOURIAN, T., BARZEGAR-JALALI, M., HAKIMIHA, N. & CRONIN, M. T. 2004. Quantitative structure-pharmacokinetic relationship modelling: apparent volume of distribution. Journal of Pharmacy and Pharmacology, 56**,** 339-350.

GIACOMINI, K. M., HUANG, S.-M., TWEEDIE, D. J., BENET, L. Z., BROUWER, K. L. R., CHU, X., DAHLIN, A., EVERS, R., FISCHER, V., HILLGREN, K. M., HOFFMASTER, K. A., ISHIKAWA, T., KEPPLER, D., KIM, R. B., LEE, C. A., NIEMI, M., POLLI, J. W., SUGIYAMA, Y., SWAAN, P. W., WARE, J. A., WRIGHT, S. H., YEE, S. W., ZAMEK-GLISZCZYNSKI, M. J. & ZHANG, L. 2010. Membrane transporters in drug development. Nature Reviews Drug Discovery, 9**,** 215-36.

GIBAJA, E. & VENTURA, S. 2014. Multi-label learning: a review of the state of the art and ongoing research. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 4**,** 411-444.

GIBAJA, E. & VENTURA, S. 2015. A Tutorial on Multilabel Learning. ACM Computing Surveys, 47**,** 1-38.

GLEESON, M. P. 2008. Generation of a Set of Simple, Interpretable ADMET Rules of Thumb. Journal of Medicinal Chemistry, 51**,** 817-834.

GLEESON, P. M., HERSEY, A. & HANNONGBUA, S. 2011. In-Silico ADME Models: A General Assessment of their Utility in Drug Discovery Applications. Current Topics in Medicinal Chemistry, 11**,** 358-381.

GOLDBERG, D. E. 1989. Genetic Algorithms in Search, Optimization and Machine Learning (1st ed.). Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

GOMBAR, V. K. & HALL, S. D. 2013. Quantitative Structure–Activity Relationship Models of Clinical Pharmacokinetics: Clearance and Volume of Distribution. Journal of Chemical Information and Modeling, 53**,** 948-957.

GOMBAR, V. K., POLLI, J. W., HUMPHREYS, J. E., WRING, S. A. & SERABJIT-SINGH, C. S. 2004. Predicting P-glycoprotein substrates by a quantitative structure–activity relationship model. Journal of Chemical Information and Modeling, 93**,** 957-968.

GOODARZI, M., HEYDEN, Y. V. & FUNAR-TIMOFEI, S. 2013. Towards better understanding of feature-selection or reduction techniques for Quantitative Structure–Activity Relationship models. TrAC Trends in Analytical Chemistry, 42**,** 49-63.

GORACCI, L., CECCARELLI, M., BONELLI, D. & CRUCIANI, G. 2013. Modeling Phospholipidosis Induction: Reliability and Warnings. Journal of Chemical Information and Modeling, 53**,** 1436-1446.

GROVER, A. & BENET, L. Z. 2009. Effects of Drug Transporters on Volume of Distribution. The AAPS Journal, 11**,** 250-261.

GUPTA, R. R., GIFFORD, E. M., LISTON, T., WALLER, C. L., HOHMAN, M., BUNIN, B. A. & EKINS, S. 2010. Using Open Source Computational Tools for Predicting Human Metabolic Stability and Additional Absorption, Distribution, Metabolism, Excretion, and Toxicity Properties. Drug Metabolism and Disposition, 38**,** 2083-2090.

HAGENBUCH, B. & STIEGER, B. 2013. The SLCO (former SLC21) superfamily of transporters. Molecular Aspects of Medicine, 34**,** 396-412.

HALL, M., FRANK, E., HOLMES, G., PFAHRINGER, B., REUTEMANN, P. & WITTEN, I. H. 2009. The WEKA Data Mining Software: An Update. SIGKDD Explorations, 11.

HALL, M. A. & HOLMES, G. 2003. Benchmarking attribute selection techniques for discrete class data mining. Knowledge and Data Engineering, IEEE Transactions on, 15**,** 1437-1447.

HALL, M. A. & SMITH, L. A. Feature Selection for Machine Learning Comparing a Correlation-based Filter Approach to the Wrapper. Twelfth International FLAIRS Conference, 1999. AAAI.

HANUMEGOWDA, U. M., WENKE, G., REGUEIRO-REN, A., YORDANOVA, R., CORRADI, J. P. & ADAMS, S. P. 2010. Phospholipidosis as a Function of Basicity, Lipophilicity, and Volume of Distribution of Compounds. Chemical Research in Toxicology, 23**,** 749-755.

HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer.

HAZAI, E., HAZAI, I., RAGUENEAU-MAJLESSI, I., CHUNG, S. P., BIKADI, Z. & MAO, Q. 2013. Predicting substrates of the human breast cancer resistance protein using a support vector machine method. BMC Bioinformatics, 14.

HAZEWINKEL, M. (ed.) 1993. Encyclopaedia of Mathematics: Stochastic Approximation — Zygmund Class of Functions: Springer Netherlands.

HO, R. H. & KIM, R. B. 2014. Solute Carriers. In: RUDEK, A. M., CHAU, H. C., FIGG, D. W. & MCLEOD, L. H. (eds.) Handbook of Anticancer Pharmacokinetics and Pharmacodynamics. New York, NY: Springer New York.

HÖGLUND, P. J., NORDSTRÖM, K. J. V., SCHIÖTH, H. B. & FREDRIKSSON, R. 2011. The Solute Carrier Families Have a Remarkably Long Evolutionary History with the Majority of the Human Families Present before Divergence of Bilaterian Species. Molecular Biology and Evolution, 28**,** 1531-1541.

HOLFORD, N. & YIM, D.-S. 2016. Volume of Distribution. Translational and Clinical Pharmacology, 24**,** 74-77.

HOLLÓSY, F., VALKÓ, K., HERSEY, A., NUNHUCK, S., KÉRI, G. & BEVAN, C. 2006. Estimation of volume of distribution in humans from high throughput HPLC-based measurements of human serum albumin binding and immobilized artificial membrane partitioning. Journal of Medicinal Chemistry, 49**,** 6958-71.

HONORIO, K. M., MODA, T. L. & ANDRICOPULO, A. D. 2013. Pharmacokinetic Properties and In Silico ADME Modeling in Drug Discovery. Medicinal Chemistry, 9**,** 163-176.

HOP, C. E. C. A. 2015. Compound Attrition at the Preclinical Phase. In: ALEX, A., HARRIS, C. J. & SMITH, D. A. (eds.) Attrition in the Pharmaceutical Industry. John Wiley & Sons, Inc.

HORN, P. S. & PESCE, A. J. 2006. Reference intervals (ranges): distribution-free methods vs. normal theory. In: BUNCHER, C. R. & TSAY, J.-Y. (eds.) Statistics In the Pharmaceutical Industry. Chapman & Francis Group.

HUANG, J., MA, G., MUHAMMAD, I. & CHENG, Y. 2007. Identifying P-Glycoprotein Substrates Using a Support Vector Machine Optimized by a Particle Swarm. Journal of Chemical Information and Modeling, 47**,** 1638-1647.

ISHIKAWA, T., FUKAMI, T., NAGAKURA, M. & HIRANO, H. 2016. Current Status and Implications of Transporters: QSAR Analysis Method to Evaluate Drug-Drug Interactions of Human Bile Salt Export Pump (ABCB11/BSEP) and Prediction of Intrahepatic Cholestasis Risk. New Horizons in Predictive Drug Metabolism and Pharmacokinetics. The Royal Society of Chemistry.

IYER, P., STUMPFE, D., VOGT, M., BAJORATH, J. & MAGGIORA, G. M. 2013. Activity Landscapes, Information Theory, and Structure - Activity Relationships. Molecular Informatics, 32**,** 421-430.

JAMES, G., WITTEN, D., HASTIE, T. & TIBSHIRANI, R. 2013. An Introduction to Statistical Learning: with Applications in R, New York, NY, Springer New York.

JAPKOWICZ, N. & SHAH, M. 2011. Evaluating Learning Algorithms: A Classification Perspective. Cambridge University Press.

JAWORSKA, J., NIKOLOVA-JELIAZKOVA, N. & ALDENBERG, T. 2005. QSAR applicabilty domain estimation by projection of the training set descriptor space: a review. Alternatives to Laboratory Animals.

KANEKO, H. & FUNATSU, K. 2014. Applicability domain based on ensemble learning in classification and regression analyses. Journal of Chemical Information and Modeling, 54**,** 2469-82.

KANEKO, H. & FUNATSU, K. 2017. Applicability Domains and Consistent Structure Generation. Molecular Informatics, 36.

KANTARDZIC, M. 2011. Data Reduction. Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons, Inc.

KARLGREN, M. & BERGSTROM, C. A. S. 2016. How Physicochemical Properties of Drugs Affect Their Metabolism and Clearance. New Horizons in Predictive Drug Metabolism and Pharmacokinetics. The Royal Society of Chemistry.

KENNEDY, T. 1997. Managing the drug discovery/development interface. Drug Discovery Today, 2**,** 436-444.

KHARKAR, P. S. 2010. Two-Dimensional (2D) In Silico Models for Absorption, Distribution, Metabolism, Excretion and Toxicity (ADME/T) in Drug Discovery. Current Topics in Medicinal Chemistry, 10**,** 116-126.

KOHAVI, R., & JOHN, G. H. 1997. Wrappers for feature subset selection, Artificial Intelligence, 97, 273-324.

KONONENKO, I., ROBNIK-SIKONJA, M. & POMPE, S. U. 1996. ReliefF for estimation and discretization of attributes in classification, regression, and ILP problems. In: RAMSEY, A. (ed.) AIMSA-96. Sozopol, Bulgaria: IOS Press.

KORZEKWA, K. & NAGAR, S. 2017. Drug Distribution Part 2. Predicting Volume of Distribution from Plasma Protein Binding and Membrane Partitioning. Pharmaceutical Research, 34**,** 544–551.

KOTSIANTIS, S. B. 2013. Decision trees: a recent overview. Artificial Intelligence Review, 39**,** 261-283.

KREIN, M., HUANG, T.-W., MORKOWCHUK, L., AGRAFIOTIS, D. K. & BRENEMAN, C. M. 2012. Developing Best Practices for Descriptor-Based Property Prediction: Appropriate Matching of Datasets, Descriptors, Methods, and Expectations. Statistical Modelling of Molecular Descriptors in QSAR/QSPR. Wiley-VCH Verlag GmbH & Co. KGaA.

KUSUHARA, H., YOSHIDA, K. & SUGIYAMA, Y. 2013. In Vivo Characterization of Interactions on Transporters. In: SUGIYAMA, Y. & STEFFANSEN, B. (eds.) Transporters in Drug Development: Discovery, Optimization, Clinical Study and Regulation. New York, NY: Springer New York.

LABUTE, P., KOSSNER, M., AJAMIAN, A., SANTAVY, M. & LIN, A. Pharmacophore annotation using extended Hückel theory. 9th German Conference on Chemoinformatics, 2014. Springer, P54.

LAI, Y. 2013a. 1 - Membrane transporters and the diseases corresponding to functional defects. Transporters in Drug Discovery and Development. Woodhead Publishing.

LAI, Y. 2013b. 2 - P-glycoprotein (P-gp/MDR1)/ABCB1. Transporters in Drug Discovery and Development. Woodhead Publishing.

LAI, Y. 2013c. 3 - Multidrug resistance-associated protein 2 (MRP2/ABCC2). Transporters in Drug Discovery and Development. Woodhead Publishing.

LAI, Y. 2013d. 4 - Breast cancer resistance protein (BCRP)/ABCG2. Transporters in Drug Discovery and Development. Woodhead Publishing.

LAI, Y. 2013e. 6 - Organic anion-transporting polypeptides (OATPs/SLCOs). Transporters in Drug Discovery and Development. Woodhead Publishing.

LAI, Y. 2013f. 7 - Organic anion, organic cation and zwitterion transporters of the SLC22 and SLC47 superfamily (OATs, OCTs, OCTNs and MATEs). Transporters in Drug Discovery and Development. Woodhead Publishing.

LEE, W., GLAESER, H., SMITH, L., ROBERTS, R., MOECKEL, G., GERVASINI, G., LEAKE, B. & KIM, R. 2005. Polymorphisms in human organic anion-transporting polypeptide 1A2 (OATP1A2): implications for altered drug disposition and central nervous system drug entry. Journal of Biological Chemistry, 280**,** 9610-7.

LEGRAND, O., SIMONIN, G., BEAUCHAMP-NICOUD, A., ZITTOUN, R. & MARIE, J.-P. 1999. Simultaneous Activity of MRP1 and Pgp Is Correlated With In Vitro Resistance to

Daunorubicin and With In Vivo Resistance in Adult Acute Myeloid Leukemia. Blood, 94**,** 1046-1056.

LIANG, Y., LI, S. & CHEN, L. 2015. The physiological role of drug transporters. Protein & Cell, 6**,** 334-350.

LIU, H., MOTODA, H., SETIONO, R. & ZHAO, Z. Feature Selection: An Ever Evolving Frontier in Data Mining. In: LIU, H., MOTODA, H., SETIONO, R. & ZHAO, Z., eds. 4th International Workshop on Feature Selection in Data Mining, 2010. 4-13.

LOCHER, K. P. 2016. Mechanistic diversity in ATP-binding cassette (ABC) transporters. Nat Struct Mol Biol, 23**,** 487-493.

LOGAN, R., KONG, A. & KRISE, J. P. 2013. Evaluating the Roles of Autophagy and Lysosomal Trafficking Defects in Intracellular Distribution-Based Drug–Drug Interactions Involving Lysosomes. Journal of Pharmaceutical Sciences, 102**,** 4173-4180.

LOMBARDO, F. & JING, Y. 2016. In Silico Prediction of Volume of Distribution in Humans. Extensive Data Set and the Exploration of Linear and Nonlinear Methods Coupled with Molecular Interaction Fields Descriptors. Journal of Chemical Information and Modeling, In Press.

LOUIS, B. & AGRAWAL, V. K. 2014. Prediction of human volume of distribution values for drugs using linear and nonlinear quantitative structure pharmacokinetic relationship models. Interdisciplinary Sciences: Computational Life Sciences, 6**,** 71-83.

LOWE, R., MUSSA, H. Y., NIGSCH, F., GLEN, R. C. & MITCHELL, J. B. 2012. Predicting the mechanism of phospholipidosis. Journal of Cheminformatics, 4**,** 2.

LU-EMERSON, C., NORDEN, A. D., DRAPPATZ, J., QUANT, E. C., BEROUKHIM, R., CIAMPA, A. S., DOHERTY, L. M., LAFRANKIE, D. C., RULAND, S. & WEN, P. Y. 2011. Retrospective study of dasatinib for recurrent glioblastoma after bevacizumab failure. Journal of Neuro-oncology, 104**,** 287-91.

LU, J., R, G. M., GRULKE, C. M., CHANG, D. T., BROOKS, R. D., LEONARD, J. A., PHILLIPS, M. B., HYPES, E. D., FAIR, M. J., TORNERO-VELEZ, R., JOHNSON, J., DARY, C. C. & TAN, Y. M. 2016. Developing a Physiologically-Based Pharmacokinetic Model Knowledgebase in Support of Provisional Model Construction. PLOS Computational Biology, 12**,** e1004495.

LUACES, O., DÍEZ, J., BARRANQUERO, J., DEL COZ, J. & BAHAMONDE, A. 2012. Binary relevance efficacy for multilabel classification. Progress in Artificial Intelligence, 1**,** 303-313.

MAATEN, L. V. D. & HINTON, G. 2008. Visualizing Data using t-SNE. Journal of Machine Learning Research, 9**,** 2579-2605.

MADAN, A. K., BAJAJ, S. & DUREJA, H. 2013. Classification Models for Safe Drug Molecules. In: REISFELD, B. & MAYENO, A. N. (eds.) Computational Toxicology: Volume II. Totowa, NJ: Humana Press.

MADJAROV, G., KOCEV, D., GJORGJEVIKJ, D. & DŽEROSKI, S. 2012. An extensive experimental comparison of methods for multi-label learning. Pattern Recognition, 45**,** 3084-3104.

MAGGIORA, G. M. 2006. On Outliers and Activity Cliffs s Why QSAR Often Disappoints. Journal of Chemical Information and Modeling, 46**,** 1535.

MAK, L., MARCUS, D., HOWLETT, A., YAROVA, G., DUCHATEAU, G., KLAFFKE, W., BENDER, A. & GLEN, R. 2015. Metrabase: a cheminformatics and bioinformatics database for small molecule transporter data analysis and (Q)SAR modeling. Journal of Cheminformatics, 7**,** 31.

MANTOVANI, R. G., ROSSI, A. L. D., VANSCHOREN, J., BISCHL, B. & CARVALHO, A. C. P. L. F. 2017. Effectiveness of Random Search in SVM hyper-parameter tuning. 2015 International Joint Conference on Neural Networks.

MARQUEZ, B. & BAMBEKE, V. 2011. ABC Multidrug Transporters: Target for Pharmacokinetics and Drug-Drug Interactions Modulation of Drug. Current Drug Targets, 12**,** 600-620.

MATHEA, M., KLINGSPOHN, W. & BAUMANN, K. 2016. Chemoinformatic Classification Methods and their Applicability Domain. molecular Informatics, 35**,** 160-180.

MATSSON, P., PEDERSEN, J., NORINDER, U., BERGSTRÖM, C. S. & ARTURSSON, P. 2009. Identification of Novel Specific and General Inhibitors of the Three Major Human ATP-Binding Cassette Transporters P-gp, BCRP and MRP2 Among Registered Drugs. Pharmaceutical Research, 26**,** 1816-1831.

MITTAL, R. R., HARRIS, L., MCKINNON, R. A. & SORICH, M. J. 2009. Partial Charge Calculation Method Affects CoMFA QSAR Prediction Accuracy. Journal of Chemical Information and Modeling, 49**,** 704-709.

MONTANARI, F., ZDRAZIL, B., DIGLES, D. & ECKER, G. F. 2016. Selectivity profiling of BCRP versus P-gp inhibition: from automated collection of polypharmacology data to multi-label learning. Journal of Cheminformatics, 8**,** 7.

MUEHLBACHER, M., TRIPAL, P., ROAS, F. & KORNHUBER, J. 2012. Identification of Drugs Inducing Phospholipidosis by Novel in vitro Data. ChemMedChem, 7**,** 1925-1934.

NETZEVA, T. I., WORTH, A. P., ALDENBERG, T., BENIGNI, R., MARK, T. D., GRAMATICA, P., JAWORSKA, J. S., KAHN, S., KLOPMAN, G., CAROL, A., MYATT, G., NIKOLOVA-JELIAZKOVA, N., PATLEWICZ, G. Y. & PERKINS, R. 2005. Current Status of Methods for Defining the Applicability Domain of (Quantitative) Structure – Activity Relationships. Alternatives to Laboratory Animals, 32**,** 1-19.

NEWBY, D., FREITAS, A. A. & GHAFOURIAN, T. 2013. Coping with Unbalanced Class Data Sets in Oral Absorption Models. Journal of Chemical Information and Modeling, 53**,** 461-474.

OBACH, R. S., LOMBARDO, F. & WATERS, N. J. 2008. Trend Analysis of a Database of Intravenous Pharmacokinetic Parameters in Humans for 670 Drug Compounds. Drug Metabolism and Disposition, 36**,** 1385-1405.

OBEROI, R. K., MITTAPALLI, R. K. & ELMQUIST, W. F. 2013. Pharmacokinetic Assessment of Efflux Transport in Sunitinib Distribution to the Brain. Journal of Pharmacology and Experimental Therapeutics, 347**,** 755-764.

ONG, C. T., BABALOLA, C. P., NIGHTINGALE, C. H. & NICOLAU, D. P. 2005. Penetration, efflux and intracellular activity of tigecycline in human polymorphonuclear neutrophils (PMNs). The Journal of Antimicrobial Chemotherapy, 56**,** 498-501.

OROGO, A. M., CHOI, S. S., MINNIER, B. L. & KRUHLAK, N. L. 2012. Construction and Consensus Performance of (Q)SAR Models for Predicting Phospholipidosis Using a Dataset of 743 Compounds. Molecular Informatics, 31**,** 725-739.

OSE, A., TOSHIMOTO, K., IKEDA, K., MAEDA, K., YOSHIDA, S., YAMASHITA, F., HASHIDA, M., ISHIDA, T., AKIYAMA, Y. & SUGIYAMA, Y. 2016. Development of a Support Vector Machine-Based System to Predict Whether a Compound Is a Substrate of a Given Drug Transporter Using Its Chemical Structure. Journal of Pharmaceutical Sciences, 105**,** 2222–2230.

PAIXÃO, P., ANICETO, N., GOUVEIS, L. F., MORAIS, J. A. G. 2014. Prediction of Drug Distribution in Rat and Humans Using an Artificial Neural Networks Ensemble and a PBPK Model. Pharmaceutical research, 31, 3313-3322.

PEAKMAN, M.-C., TROUTMAN, M., GONZALES, R. & SCHMIDT, A. 2015. Experimental Screening Strategies to Reduce Attrition Risk. In: ALEX, A., HARRIS, C. J. & SMITH, D. A. (eds.) Attrition in the Pharmaceutical Industry. John Wiley & Sons, Inc.

PINTO, M., DIGLES, D. & ECKER, G. F. 2014. Computational models for predicting the interaction with ABC transporters. Drug Discovery Today: Technologies, 12**,** e69-77.

PINTO, M., TRAUNER, M. & ECKER, G. F. 2012. An In Silico Classification Model for Putative ABCC2 Substrates. Molecular informatics, 31**,** 547-553.

POLISHCHUK, P., TINKOV, O., KHRISTOVA, T., OGNICHENKO, L., KOSINSKAYA, A., VARNEK, A. & KUZ'MIN, V. 2016. Structural and Physico-Chemical Interpretation (SPCI) of QSAR Models and Its Comparison with Matched Molecular Pair Analysis. Journal of chemical information and modeling, 56**,** 1455-1469.

PRENTIS, R. A., LIS, Y. & WALKER, S. R. 1988. Pharmaceutical innovation by the seven UK-owned pharmaceutical companies (1964-1985). British Journal of Clinical Pharmacology, 25**,** 387-396.

QUINLAN, J. R. 1993. C4.5: programs for machine learning, Morgan Kaufmann Publishers Inc.

READ, J., PFAHRINGER, B., HOLMES, G. & FRANK, E. 2009. Classifier Chains for Multi-label Classification. In: BUNTINE, W., GROBELNIK, M., MLADENIĆ, D. & SHAWE-TAYLOR, J. (eds.) Machine Learning and Knowledge Discovery in Databases. Springer Berlin Heidelberg.

REASOR, M. J., HASTINGS, K. L. & ULRICH, R. G. 2006. Drug-induced phospholipidosis: issues and future directions. Expert Opinion on Drug Safety, 5**,** 567-583.

RECANATINI, M. & CAVALLI, A. 2008. QSAR and Pharmacophores for Drugs Involved in hERG Blockage. Antitargets. Wiley-VCH Verlag GmbH & Co. KGaA.

ROBNIK-ŠIKONJA, M. & KONONENKO, I. 2003. Theoretical and Empirical Analysis of ReliefF and RReliefF. Machine Learning, 53**,** 23-69.

ROKACH, L. & MAIMON O. 2015. Data Mining With Decision Trees: Theory And Applications (2nd Edition). World Scientific Publishing Co., Inc., River Edge, NJ, USA.

ROTH, M., OBAIDAT, A. & HAGENBUCH, B. 2012. OATPs, OATs and OCTs: the organic anion and cation transporters of the SLCO and SLC22A gene superfamilies. British Journal Pharmacology, 165**,** 1260-87.

ROY, K., KAR, S. & DAS, R. N. 2015. Statistical Methods in QSAR/QSPR. A Primer on QSAR/QSPR Modeling: Fundamental Concepts. Cham: Springer International Publishing.

SAEYS, Y., INZA, I. & LARRAÑAGA, P. 2007. A review of feature selection techniques in bioinformatics. Bioinformatics, 23**,** 2507-2517.

SAHIGARA, F., BALLABIO, D., TODESCHINI, R. & CONSONNI, V. 2013. Defining a novel k-nearest neighbours approach to assess the applicability domain of a QSAR model for reliable predictions. Journal of Cheminformatics, 5**,** 27.

SAHIGARA, F., MANSOURI, K., BALLABIO, D., MAURI, A., CONSONNI, V. & TODESCHINI, R. 2012. Comparison of different approaches to define the applicability domain of QSAR models. Molecules, 17**,** 4791-810.

SAHLIN, U. 2013. Uncertainty in QSAR Predictions. Alternatives to Laboratory Animals, 41**,** 111-125.

SAHLIN, U., JELIAZKOVA, N. & ÖBERG, T. 2014. Applicability Domain Dependent Predictive Uncertainty in QSAR Regressions. Molecular Informatics, 33**,** 26-35.

SCANNELL, J. W., BLANCKLEY, A., BOLDON, H. & WARRINGTON, B. 2012. Diagnosing the decline in pharmaceutical R&D efficiency. Nature Reviews Drug Discovery, 11**,** 191-200.

SCHLESSINGER, A., KHURI, N., GIACOMINI, K. M. & SALI, A. 2013a. Molecular modeling and ligand docking for solute carrier (SLC) transporters. Current Topics in Medicinal Chemistry, 13**,** 843-56.

SCHLESSINGER, A., MATSSON, P., SHIMA, J. E., PIEPER, U., YEE, S. W., KELLY, L., APELTSIN, L., STROUD, R. M., FERRIN, T. E., GIACOMINI, K. M. & SALI, A. 2010. Comparison of human solute carriers. Protein Science, 19**,** 412-28.

SCHLESSINGER, A., YEE, S. W., SALI, A. & GIACOMINI, K. M. 2013b. SLC Classification: An Update. Clinical Pharmacology & Therapeutics, 94**,** 19-23.

SECHIDIS, K., TSOUMAKAS, G. & VLAHAVAS, I. 2011. On the Stratification of Multi-Label Data. In: GUNOPULOS, D., HOFMANN, T., MALERBA, D. & VAZIRGIANNIS, M. (eds.) ECML PKDD 2011. Greece: Springer.

SEDYKH, A., FOURCHES, D., DUAN, J., HUCKE, O., GARNEAU, M., ZHU, H., BONNEAU, P. & TROPSHA, A. 2013. Human intestinal transporter database: QSAR modeling and virtual profiling of drug uptake, efflux and interactions. Pharmaceutical Research, 30**,** 996-1007.

SHAHLAEI, M. 2013. Descriptor Selection Methods in Quantitative Structure–Activity Relationship Studies: A Review Study. Chemical Reviews, 113**,** 8093-8103.

SHAIKH, N., SHARMA, M. & GARG, P. 2017. Selective Fusion of Heterogeneous Classifiers for Predicting Substrates of Membrane Transporters. Journal of Chemical Information and Modeling, 57**,** 594–607.

SHARIFI, M. & GHAFOURIAN, T. 2016. Effect of OATP-binding on the prediction of biliary excretion. Xenobiotica**,** 1-18.

SHAYMAN, J. A. & ABE, A. 2013. Drug induced phospholipidosis: An acquired lysosomal storage disorder. Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids, 1831**,** 602-611.

SHERIDAN, R. P. 2012. Three useful dimensions for domain applicability in QSAR models using random forest. Journal of Chemical Information and Modeling, 52**,** 814-23.

SHERIDAN, R. P., FEUSTON, B. P., MAIOROV, V. N. & KEARSLEY, S. K. 2004. Similarity to Molecules in the Training Set Is a Good Discriminator for Prediction Accuracy in QSAR. Journal of Chemical Information and Computer Sciences, 44**,** 1912-1928.

SHIMAZAKI, H. & SHINOMOTO, S. 2010. Kernel bandwidth optimization in spike rate estimation. Journal of Computational Neuroscience, 29**,** 171-182.

SHIRASAKA, Y., MORI, T., SHICHIRI, M., NAKANISHI, T. & TAMAI, I. 2012. Functional Pleiotropy of Organic Anion Transporting Polypeptide OATP2B1 Due to Multiple Binding Sites. Drug Metabolism and Pharmacokinetics, 27**,** 360-364.

SMITH, D. A. 2016. Physicochemistry and the Off-Target Effects of Drug Molecules. Drug Discovery Toxicology. John Wiley & Sons, Inc.

SMITH, D. A. & BAILLIE, T. A. 2015. Attrition in Phase I. Attrition in the Pharmaceutical Industry. John Wiley & Sons, Inc.

SMITH, D. A., BEAUMONT, K., MAURER, T. S. & DI, L. 2015. Volume of Distribution in Drug Design. Journal of Medicinal Chemistry, 58**,** 5691-5698.

SPOLAÔR, N., CHERMAN, E. A., MONARD, M. C. & LEE, H. D. 2013. A Comparison of Multi-label Feature Selection Methods using the Problem Transformation Approach. Electron Notes Theor Comput Sci, 292**,** 135-151.

STROBL, C., BOULESTEIX, A.-L., ZEILEIS, A. & HOTHORN, T. 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics, 8**,** 25.

STUMPFE, D. & BAJORATH, J. 2012. Exploring Activity Cliffs in Medicinal Chemistry. Journal of Medicinal Chemistry, 55**,** 2932-2942.

SUI, X., SUN, J., LI, H., WANG, Y., LIU, J., LIU, X., ZHANG, W., CHEN, L. & HE, Z. 2009. Prediction of volume of distribution values in human using immobilized artificial membrane partitioning coefficients, the fraction of compound ionized and plasma protein binding data. European Journal of Medicinal Chemistry, 44**,** 4455-60.

SUSHKO, I., NOVOTARSKYI, S., KO, R., PANDEY, A. K., CHERKASOV, A., LIU, H., YAO, X., TOMAS, O., HORMOZDIARI, F., DAO, P., SAHINALP, C., TODESCHINI, R., POLISHCHUK, P., ARTEMENKO, A., KUZ, V., MARTIN, T. M., YOUNG, D. M., FOURCHES, D., MURATOV, E., TROPSHA, A., BASKIN, I., HORVATH, D., MARCOU, G., MULLER, C., VARNEK, A., PROKOPENKO, V. V. & TETKO, I. V. 2010a. Applicability Domains for Classification Problems : Benchmarking of Distance to Models for Ames Mutagenicity Set. Journal of Chemical Information and Modeling, 50**,** 2094-2111.

SUSHKO, I., NOVOTARSKYI, S., KÖRNER, R., PANDEY, A. K., KOVALISHYN, V. V., PROKOPENKO, V. V. & TETKO, I. V. 2010b. Applicability domain for in silico models to achieve accuracy of experimental measurements. Journal of Chemometrics, 24**,** 202-208.

SUSHKO, Y., NOVOTARSKYI, S., KÖRNER, R., VOGT, J., ABDELAZIZ, A. & TETKO, I. 2014. Prediction-driven matched molecular pairs to interpret QSARs and aid the molecular optimization process. Journal of Cheminformatics, 6**,** 1-18.

SUTHERLAND, J. J., RAYMOND, J. W., STEVENS, J. L., BAKER, T. K. & WATSON, D. E. 2012. Relating Molecular Properties and in Vitro Assay Results to in Vivo Drug Disposition and Toxicity Outcomes. Journal of Medicinal Chemistry, 55**,** 6455-6466.

SZAKÁCS, G., VÁRADI, A., OZVEGY-LACZKA, C. & SARKADI, B. 2008. The role of ABC transporters in drug absorption, distribution, metabolism, excretion and toxicity (ADME-Tox). Drug Discovery Today, 13**,** 379-93.

TAMAI, I. & NAKANISHI, T. 2013. Analysis of Intestinal Transporters. In: SUGIYAMA, Y. & STEFFANSEN, B. (eds.) Transporters in Drug Development: Discovery, Optimization, Clinical Study and Regulation. New York, NY: Springer New York.

TANG, J., ALELYANI, S. & LIU, H. 2014. Feature Selection for Classification: A Review. In: AGGARWAL, C. C. (ed.) Data Classification: Algorithms and Applications. Florida: CRC Press.

TASHIMA, T. 2015. Intriguing possibilities and beneficial aspects of transporter-conscious drug design. Bioorganic & Medicinal Chemistry, 23**,** 4119-4131.

TETKO, I. V., NOVOTARSKYI, S., SUSHKO, I., IVANOV, V., PETRENKO, A. E., DIEDEN, R., LEBON, F. & MATHIEU, B. 2013. Development of Dimethyl Sulfoxide Solubility Models Using 163 000 Molecules: Using a Domain Applicability Metric to Select More Reliable Predictions. Journal of Chemical Information and Modeling, 53**,** 1990-2000.

TETKO, I. V., SUSHKO, I., PANDEY, A. K., ZHU, H., TROPSHA, A., PAPA, E., TODESCHINI, R., FOURCHES, D. & VARNEK, A. 2008. 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, 48**,** 1733-1746.

TIAN, S., WANG, J., LI, Y., LI, D., XU, L. & HOU, T. 2015. The application of in silico drug-likeness predictions in pharmaceutical research. Advanced Drug Delivery Reviews, 86**,** 2-10.

TIRONA, R. G. & KIM, R. B. 2014. Organic Anion-Transporting Polypeptides. Drug Transporters. John Wiley & Sons, Inc.

TONG, W., XIE, Q., HONG, H., SHI, L., FANG, H. & PERKINS, R. 2004. Assessment of prediction confidence and domain extrapolation of two structure-activity relationship models for predicting estrogen receptor binding activity. Environmental Health Perspectives, 112**,** 1249-1254.

TOPLAK, M., MOČNIK, R., POLAJNAR, M., BOSNIĆ, Z., CARLSSON, L., HASSELGREN, C., DEMŠAR, J., BOYER, S., ZUPAN, B. & STÅLRING, J. 2014. Assessment of Machine Learning Reliability Methods for Quantifying the Applicability Domain of QSAR Regression Models. Journal of Chemical Information and Modeling, 54**,** 431-441.

TROPSHA, A. 2010. Best Practices for QSAR Model Development, Validation, and Exploitation. Molecular Informatics, 29**,** 476-488.

TROPSHA, A. & GOLBRAIKH, A. 2010. Predictive Quantitative Structure-Activity Relationships Modelling: Data Preparation and General Modeling workflow. In: FAULON, J.-L. & BENDER, A. (eds.) Handbook of Chemoinformatics Algorithms. Chapman & Hall/CRC.

TSAIOUN, K., BLAAUBOER, B. J. & HARTUNG, T. 2016. Evidence-based absorption, distribution, metabolism, excretion (ADME) and its interplay with alternative toxicity methods. ALTEX, 33**,** 343-358.

TSAIOUN, K. & KATES, S. A. 2012. ADME (Absorption, Distribution, Metabolism, Excretion): The Real Meaning - Avoiding Disaster and Maintaining Efficacy for Preclinical Candidates. In: LAPCHAK, P. A. & ZHANG, J. H. (eds.) Translational Stroke Research: From Target Selection to Clinical Trials. New York, NY: Springer New York.

TSOUMAKAS, G. & KATAKIS, I. 2007. Multi-Label Classification: An Overview. International Journal of Data Warehousing and Mining, 3**,** 1-13.

TSOUMAKAS, G., KATAKIS, I. & VLAHAVAS, I. 2010. Mining Multi-label Data. In: MAIMON, O. & ROKACH, L. (eds.) Data Mining and Knowledge Discovery Handbook. New York: Springer.

TU, M., MATHIOWETZ, A. M., PFEFFERKORN, J. A., CAMERON, K. O., DOW, R. L., LITCHFIELD, J., DI, L., FENG, B. & LIRAS, S. 2013. Medicinal Chemistry Design Principles for Liver Targeting Through OATP Transporters. Current Topics in Medicinal Chemistry, 13**,** 857-866.

UHLÉN, M., FAGERBERG, L., HALLSTRÖM, B. M., LINDSKOG, C., OKSVOLD, P., MARDINOGLU, A., SIVERTSSON, Å., KAMPF, C., SJÖSTEDT, E., ASPLUND, A., OLSSON, I., EDLUND, K., LUNDBERG, E., NAVANI, S., SZIGYARTO, C. A.-K., ODEBERG, J., DJUREINOVIC, D., TAKANEN, J. O., HOBER, S., ALM, T., EDQVIST, P.-H., BERLING, H., TEGEL, H., MULDER, J., ROCKBERG, J., NILSSON, P., SCHWENK, J. M., HAMSTEN, M., VON FEILITZEN, K., FORSBERG, M., PERSSON, L., JOHANSSON, F., ZWAHLEN, M., VON HEIJNE, G., NIELSEN, J. & PONTÉN, F. 2015. Tissue-based map of the human proteome. Science, 347.

VAN DE WATERBEEMD, H. & GIFFORD, E. 2003. ADMET in silico modelling: towards prediction paradise? Nat Rev Drug Discov, 2**,** 192-204.

VASTAG, M., HELLINGER, E., BAKK, M. L. & TIHANAYI, K. 2011. Cell-based models of blood – brain barrier penetration. Therapeutic Delivery, 2**,** 549-553.

WAGNER, D. J., HU, T. & WANG, J. 2016. Polyspecific organic cation transporters and their impact on drug intracellular levels and pharmacodynamics. Pharmacological Research, 111**,** 237-246.

WALLACE, G. C., RAMSDEN, D. B., GRANT, M. H. & LYUBIMOV, A. V. 2011. General Principles of Drug Distribution. Encyclopedia of Drug Metabolism and Interactions. John Wiley & Sons, Inc.

WANG, Y., XING, J., XU, Y., ZHOU, N., PENG, J., XIONG, Z., LIU, X., LUO, X., LUO, C., CHEN, K., ZHENG, M. & JIANG, H. 2015. In silico ADME/T modelling for rational drug design. Quarterly Reviews of Biophysics, 48**,** 488-515.

WANG, Z., CHEN, Y., LIANG, H., BENDER, A., GLEN, R. C. & YAN, A. 2011. P-glycoprotein Substrate Models Using Support Vector Machines Based on a Comprehensive Data set. Journal of Chemical Information and Modeling, 51**,** 1447–1456.

WARING, M. J., ARROWSMITH, J., LEACH, A. R., LEESON, P. D., MANDRELL, S., OWEN, R. M., PAIRAUDEAU, G., PENNIE, W. D., PICKETT, S. D., WANG, J., WALLACE, O. & WEIR, A. 2015. An analysis of the attrition of drug candidates from four major pharmaceutical companies. Nat Rev Drug Discov, 14**,** 475-486.

WASSERMANN, A. M., DIMOVA, D. & BAJORATH, J. 2011. Comprehensive analysis of single- and multi-target activity cliffs formed by currently available bioactive compounds. Chemical Biology & Drug Design, 78**,** 224-8.

WATERS, N. J. & LOMBARDO, F. 2010. Use of the Øie-Tozer Model in Understanding Mechanisms and Determinants of Drug Distribution. Drug Metabolism and Disposition, 38**,** 1159-1165.

WATTS, N. B. & DIAB, D. L. 2010. Long-Term Use of Bisphosphonates in Osteoporosis. The Journal of Clinical Endocrinology & Metabolism, 95**,** 1555-1565.

WEBB, G. I. 2000. MultiBoosting: A Technique for Combining Boosting and Wagging. Machine Learning, 40**,** 159-196.

WHO 2013. Control and Surveillance of Human African Trypanosomiasis. Italy: World Health Organization.

WILDMAN, S. A. & CRIPPEN, G. M. 1999. Prediction of Physicochemical Parameters by Atomic Contributions. Journal of Chemical Information and Computer Sciences, 39**,** 868-873.

WIND, N. S. & HOLEN, I. 2011. Multidrug resistance in breast cancer: from in vitro models to clinical studies. International Journal of Breast Cancer 2011**,** 967419.

WINTER, G. E., RADIC, B., MAYOR-RUIZ, C., BLOMEN, V. A., TREFZER, C., KANDASAMY, R. K., HUBER, K. V. M., GRIDLING, M., CHEN, D., KLAMPFL, T., KRALOVICS, R., KUBICEK, S., FERNANDEZ-CAPETILLO, O., BRUMMELKAMP, T. R. & SUPERTI-FURGA, G. 2014. The solute carrier SLC35F2 enables YM155-mediated DNA damage toxicity. Nature Chemical Biology, 10**,** 768-773.

WISHART, D. S. 2007. Improving Early Drug Discovery through ADME Modelling. Drugs in R & D, 8**,** 349-362.

WITTEN, I., FRANK, E. & HALL, M. A. 2011. Data Mining: Practical Machine Learning Tools and Techniques, Second Edition, Elsevier.

YANNI, S. B. 2015. Translational ADMET for Drug Therapy: Principles, Methods, and Pharmaceutical Applications, Wiley.

YARIM, M. & KOKSAL, M. 2010. Organic Anion Transporting Polypeptides (Oatps/OATPs). Transporters as Drug Carriers. Wiley-VCH Verlag GmbH & Co. KGaA.

YOUSEFINEJAD, S. & HEMMATEENEJAD, B. 2015. Chemometrics tools in QSAR/QSPR studies: A historical perspective. Chemometrics and Intelligent Laboratory Systems, 149**,** 177-204.

ZHANG, D., LUO, G., DING, X. & LU, C. 2012. Preclinical experimental models of drug metabolism and disposition in drug discovery and development. Acta Pharmaceutica Sinica B, 2**,** 549-561.

ZHANG, M. L. & ZHOU, Z. H. 2014. A review on multi-label learning algorithms. IEEE Transactions on Knowledge and Data Engineering, 26**,** 1819-1837.

ZHAO, Z., WANG, L., LIU, H. & YE, J. 2013. On Similarity Preserving Feature Selection. IEEE Transactions on Knowledge and Data Engineering, 25**,** 619-632.

ZHENG, N., ZHANG, X. & ROSANIA, G. R. 2011. Effect of Phospholipidosis on the Cellular Pharmacokinetics of Chloroquine. Journal of Pharmacology and Experimental Therapeutics, 336**,** 661-671.

ZHIVKOVA, Z. & DOYTCHINOVA, I. 2012. Prediction of Steady-State Volume of Distribution of Acidic Drugs by Quantitative Structure–Pharmacokinetics Relationships. Journal of Pharmaceutical Sciences, 101**,** 1253-1266.

ZHIVKOVA, Z. D., MANDOVA, T. & DOYTCHINOVA, I. 2015. Quantitative Structure - Pharmacokinetics Relationships Analysis of Basic Drugs: Volume of Distribution. Journal of Pharmacy & Pharmaceutical Sciences, 18**,** 515-27.

ZHONG, L., MA, C.-Y., ZHANG, H., YANG, L.-J., WAN, H.-L., XIE, Q.-Q., LI, L.-L. & YANG, S.-Y. 2011. A prediction model of substrates and non-substrates of breast cancer resistance protein (BCRP) developed by GA-CG-SVM method. Computers in Biology and Medicine, 41**,** 1006-13.

ZHOU, J., XU, J., HUANG, Z. & WANG, M. 2015. Transporter-mediated tissue targeting of therapeutic molecules in drug discovery. Bioorganic & Medicinal Chemistry Letters, 25**,** 993-997.

11. Appendices

11.1. Appendix I: Supporting Information for Chapter 4

Table A1.1. Performances of all single label models for the 5 different feature selection methods used. Performance on the training set is presented to left and on the internal validation to the right (in grey). From within each transporter the best feature selection method was chosen based solely on the best performance on the validation.

GA

	BCRP1		MDR1		MRP1		MRP2	
TP	163	45	305	97	51	11	77	17
TN	67	17	124	28	50	15	54	13
FP	41	18	123	56	4	3	11	11
FN	17	16	28	13	6	8	3	10
ACC	0.80	0.65	0.74	0.64	0.91	0.70	0.90	0.59
Sen	0.91	0.74	0.92	0.88	0.89	0.58	0.96	0.63
Spe	0.62	0.49	0.50	0.33	0.93	0.83	0.83	0.54
MCC	0.56	0.23	0.47	0.26	0.82	0.42	0.81	0.17

GS

	BCRP1		MDR1		MRP1		MRP2	
TP	170	47	217	66	51	9	74	17
TN	98	22	197	56	52	15	54	14
FP	10	13	50	28	2	3	11	10
FN	10	14	116	44	6	10	6	10
ACC	0.93	0.72	0.71	0.63	0.93	0.65	0.88	0.61
Sen	0.94	0.77	0.65	0.60	0.89	0.47	0.93	0.63
Spe	0.91	0.63	0.80	0.67	0.96	0.83	0.83	0.58
MCC	0.85	0.40	0.45	0.26	0.86	0.33	0.76	0.21

J48-GA

	BCRP1		MDR1		MRP1		MRP2	
TP	166	45	297	79	52	17	80	23
TN	73	15	231	55	52	11	43	7
FP	35	20	16	29	2	7	22	17
FN	14	16	36	31	5	2	0	4
ACC	0.83	0.63	0.91	0.69	0.94	0.76	0.85	0.59
Sen	0.92	0.74	0.89	0.72	0.91	0.89	1.00	0.85
Spe	0.68	0.43	0.94	0.65	0.96	0.61	0.66	0.29
MCC	0.63	0.17	0.82	0.37	0.88	0.53	0.72	0.17

RfF

	BCRP1		MDR1		MRP1		MRP2	
TP	162	45	291	91	51	12	71	17
TN	100	20	111	30	52	11	61	16
FP	8	15	136	54	2	7	4	8
FN	18	16	42	19	6	7	9	10
ACC	0.91	0.68	0.69	0.62	0.93	0.62	0.91	0.65
Sen	0.90	0.74	0.87	0.83	0.89	0.63	0.89	0.63
Spe	0.93	0.57	0.45	0.36	0.96	0.61	0.94	0.67
MCC	0.81	0.31	0.36	0.21	0.86	0.24	0.82	0.30

Rf-GS

	BCRP1		MDR1		MRP1		MRP2	
TP	141	48	321	86	56	15	76	18
TN	76	20	198	37	51	12	57	15
FP	32	15	49	47	3	6	8	9
FN	39	13	12	24	1	4	4	9
ACC	0.75	0.71	0.89	0.63	0.96	0.73	0.92	0.65
Sen	0.78	0.79	0.96	0.78	0.98	0.79	0.95	0.67
Spe	0.70	0.57	0.80	0.44	0.94	0.67	0.88	0.63
MCC	0.48	0.36	0.79	0.24	0.93	0.46	0.83	0.29

Table A1.2 Definitions of all molecular descriptors present in the BR and CC models.

Feature	Models where the feature is present	Definition
a_acc	MDR1-CC/BR	Number of hydrogen bond acceptor atoms (not counting acidic atoms but counting atoms that are both hydrogen bond donors and acceptors such as -OH).
a_aro	BCRP1-CC	Number of aromatic atoms
a_ICM	BCRP1-CC/ BR	The mean atom information content. The entropy of each atom's distribution in the molecule (including implicit hydrogens; excluding lone pair pseudo-atoms). a_ICM= -sum(pi.log[pi]), where pi = atom i count/total atom count.
a_nF	BCRP1-CC/ BR	Number of F atoms
a_nH	MDR1-CC/BR	Number of H atoms
ast_violation_ext	MRP2-BR/CC	Astex violations (extended), otherwise known as the rule of 3
b_ar	BCRP1-CC/ BR MRP1-CC/BR	Number of aromatic bonds
b_max1len	MDR1-CC/BR MRP1-BR/CC	Maximum length of a single bond path.

Table A1.2 (cont.)

Feature	Models where the feature is present	Definition
b_rotR	MRP1-BR/ CC MRP2-BR/ CC	Fraction of rotatable bonds: Number of rotatable bonds divided by number of bonds between heavy atoms.
chi1v_C	MRP1-BR	Carbon valence connectivity index (order 1). This is calculated as the sum of 1/sqrt(vivj) over all bonds between carbon atoms i and j where i < j. vi = (pi - hi) / (Zi - pi - 1) where pi is the number of s and p valence electrons of atom i, Z is the atomic number, and hi is the number of H ideally bound to atom i.
dens	MDR1-CC/BR	Mass density: molecular weight divided by van der Waals volume as calculated in the vol descriptor
FCASA-	MDR1-CC/BR	Fractional CASA- calculated as CASA- / ASA. Here, CASA- is the negative charge weighted surface area calculated as ASA-*max $\{q_i < 0\}$, where ASA- is water accessible surface area of negatively charged atoms. ASA is the total water accessible surface area.
FCASA+	MRP1-BR	Fractional CASA+ calculated as CASA+ / ASA. The negative equivalent of FCASA- definition.
FCharge	MDR1-CC/BR	Formal charge
Fi(A)	MRP2-BR	Fraction of ionized acid at pH 7.4.
Fi(B)	MRP2-BR/ CC	Fraction of ionized base at pH 7.4.
glob	BCRP1-CC/ BR	Globularity. The ratio of surface to the surface of a same-volume sphere.
Kier3	MRP1-BR	Third kappa shape index: $(n-1)\ (n-3)^2 / p3^2$ for odd n, and $(n-3)\ (n-2)^2 / p3^2$ for even n.
LogD(5.5)	BCRP1-CC/ BR	Log(octanol/water at pH 5.5)
LogD(6.5)	BCRP1-CC/ BR	Log(octanol/water at pH 6.5)
LogD(7.4)	BCRP1-CC/ BR	Log(octanol/water at pH 7.4)
MNDO_LUMO	BCRP1-CC/ BR	The energy (eV) of the Lowest Unoccupied Molecular Orbital calculated using the MNDO Hamiltonian
MW	MDR1-CC/BR	Molecular weight
Num_Rings_4	BCRP1-CC/ BR	Number of 4-member rings
Num_Rings_5	MDR1-CC/BR	Number of 6-member rings
opr_leadlike	MRP2-BR	One if and only if opr_violation < 2 otherwise zero.
opr_nring	BCRP1-CC/ BR	Oprea's ring count. Rings are counted according to the
PEOE_VSA_FPNEG	MRP2-BR/ CC	Fractional negative polar van der Waals surface area. The sum of VSA for atoms with atomic charge < -0.2, divided by the total surface area.

Table A1.2 (cont.)

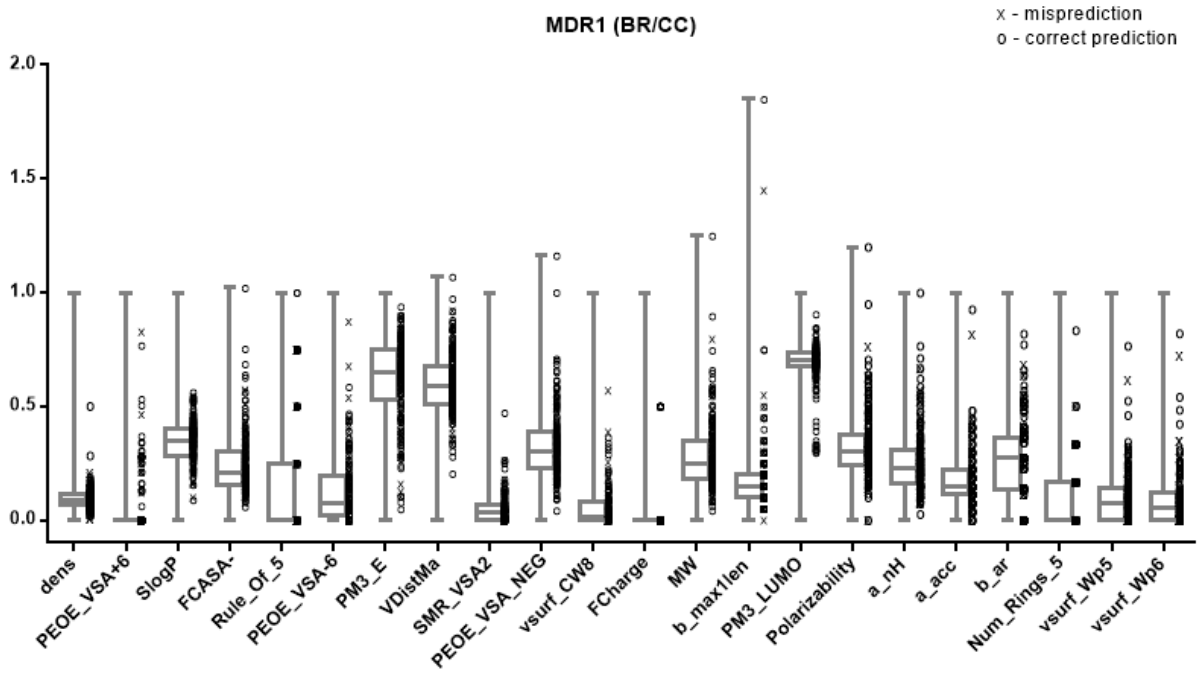
Feature	Models where the feature is present	Definition
PEOE_VSA_NEG	MDR1-CC/BR	Total negative van der Waals surface area. The sum VDW contribution from atoms with negative atomic charge.
PEOE_VSA+6	MDR1-CC/BR	Sum of VDW volume of atoms where the atomic charge greater than 0.3
PEOE_VSA-1	BCRP1-BR	Sum of VDW volume of atoms where the atomic charge is in the range [-0.25,-0.20).
PEOE_VSA-3	BCRP1-CC/ BR	Sum of VDW volume of atoms where the atomic charge is in the range [-0.20,-0.15)
PEOE_VSA-4	BCRP1-CC/BR	Sum of VDW volume of atoms where the atomic charge is in the range [-0.10,-0.05).
PEOE_VSA-6	MDR1-CC/BR	Sum of VSA of atoms with atomic charge > -0.3
PM3_E	MDR1-CC/BR	The total self-consistent field energy (kcal/mol) calculated using the PM3 Hamiltonian.
PM3_LUMO	MDR1-CC/BR	The energy (eV) of the Lowest Unoccupied Molecular Orbital calculated using the PM3 Hamiltonian.
pMDR1_J48-GA	BCRP1-CC	Predictions from the MRP1-CC single-label model
pMRP2_RfF	MRP1-CC	Predictions from the MRP2-CC single-label model
Polarizability	MDR1-CC/BR	Propensity for formation of momentary dipoles upon interaction with an electrically charge species.
Q_VSA_FHYD	MRP2-BR/ CC	Fractional hydrophobic van der Waals surface area. This is the sum of VSA of atoms with partial charge less than or equal to 0.2 divided by the total surface area.
Q_VSA_FPPOS	BCRP1-CC/ BR MRP1-CC/ BR	Fraction of positive polar VSA (PM6-derived) defined as the van der Waals area of atoms with atomic charge > 0.2, divided by the total surface are
Q_VSA_POL	MRP1-BR/ CC	Total polar van der Waals surface area. The sum of VDW surface of atoms with absolute partial charge greater than 0.2.
reactive	MRP2-BR/ CC	Reactive groups count. The table of reactive groups is based on the Oprea set and includes metals, phospho-, N/O/S-N/O/S single bonds, thiols, acyl halides, Michael Acceptors, azides, esters, etc.
rings	MRP1-CC	The number of rings.
Rule_Of_5	MDR1-CC/BR	Rule of Five
SlogP	MDR1-CC/BR	Log of the octanol/water partition coefficient calculated from the sum of individual contributions to logP from each atom.*
SlogP_VSA7	BCRP1-CC/BR	VSA of atoms which contributing logP is (0.25,0.30].*
SMR_VSA2	MDR1-CC/BR	Total VDW surface area of atoms with molecular refractivity ranging (0.26,0.35]*

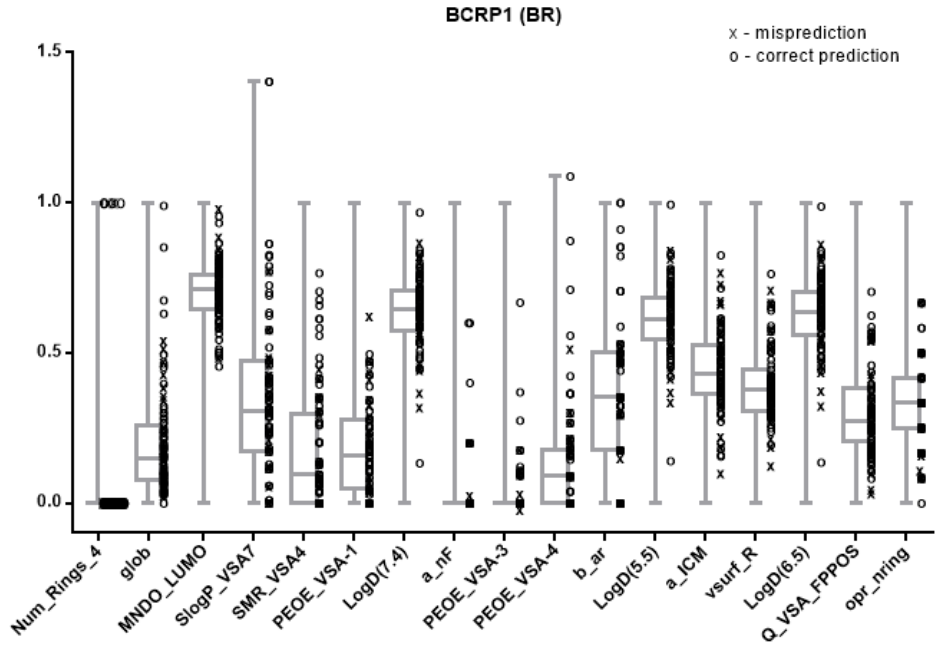
Table A1.2 (cont.)

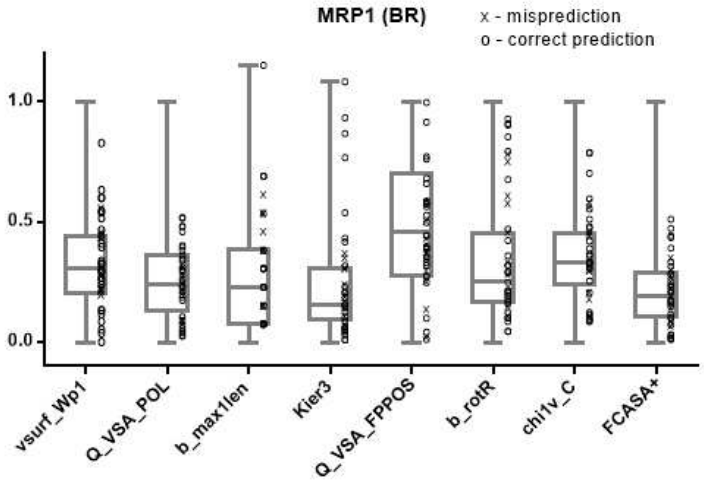
Feature	Models where the feature is present	Definition
SMR_VSA4	BCRP1-CC/BR	Total VDW surface area of atoms with molecular refractivity ranging (0.39,0.44]*
VDistMa	MDR1-CC/BR	VDistMa is defined to be the sum of $\log_2 m - D_{ij} \log_2 D_{ij} / m$ over all i and j. D is a distance matrix between every atom i and j; m is the sum of the distance matrix entries.
vsurf_CW2	MRP2-CC/BR BCRP1-BR	Capacity factor, calculated as the ratio of the hydrophilic surface to the total molecular surface, at -0.5kcal/mol.
vsurf_CW8	MDR1-CC/BR	Capacity factor, calculated as the ratio of the hydrophilic surface to the total molecular surface, at -6kcal/mol.
vsurf_D7	MRP1-CC	Volume of the hydrophobic interactions at -1.4 kcal/mol
vsurf_HB2	MRP2-BR/ CC	Hydrogen-bond donor capacity. Defined as the difference between the volume of the hydrophilic interactions vsurf_W2 and volume of the O probe interactions vsurf_Wp2.
vsurf_R	BCRP1-CC/ BR	Surface rugosity, defined as the ratio of volume to surface
vsurf_Wp1	MRP1-BR	Polar volume. Volume of the interactions with an O probe at -0.2kcal/mol
vsurf_Wp5	MDR1-CC/BR	Polar volume. Volume of the interactions with an O probe at -3kcal/mol
vsurf_Wp6	MDR1-CC/BR	Polar volume. Volume of the interactions with an O probe at -4kcal/mol

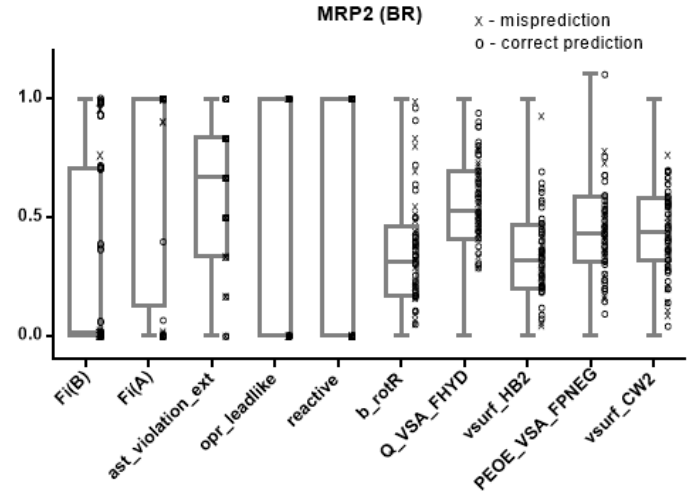
*Atom contribution values listed on Wildman and Crippen (Wildman and Crippen, 1999).

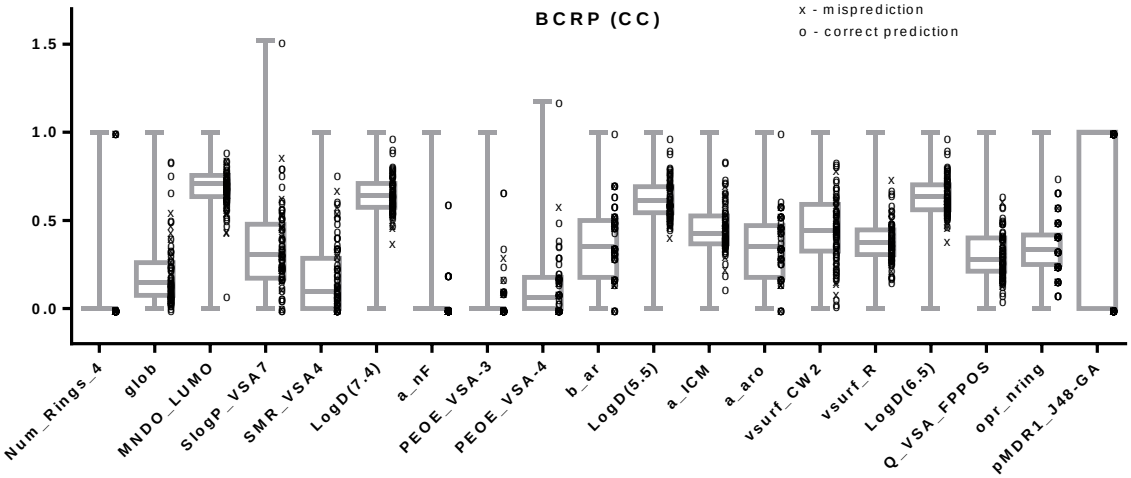
Figures A1.1-7. Misprediction analysis. Distribution of test set compounds with mis-predicted cases highlighted across the training span of each feature used in the various models.

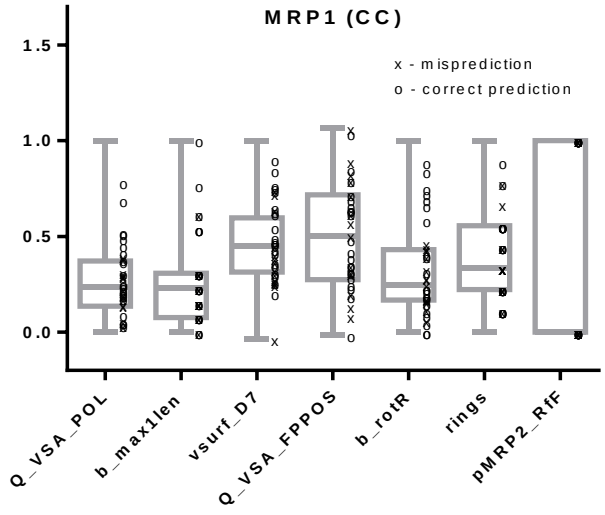


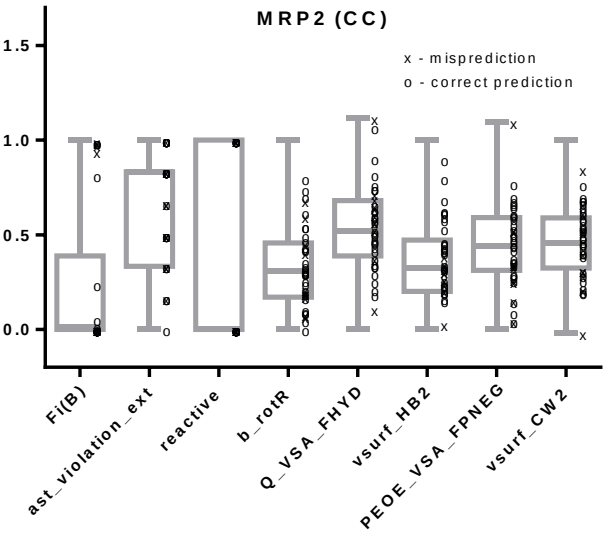



MRP1 (BR)
x - misprediction
o - correct prediction
1.0
0.5
0.0
vsurf_Wp1
Q_VSA_POL
b_max1len
Kier3
Q_VSA_FPPOS
b_rotR
chi1v_C
FCASA+


MRP2 (BR)
x - misprediction
o - correct prediction
1.0
0.5
0.0
Fi(B)
Fi(A)
ast_violation_ext
opr_leadlike
reactive
b_rotR
Q_VSA_FHYD
vsurf_HB2
PEOE_VSA_FPNEG
vsurf_CW2


BCRP (CC)
x - misprediction
o - correct prediction
1.5
1.0
0.5
0.0
Num_Rings_4
glob
MNDO_LUMO
SlogP_VSA7
SMR_VSA4
LogD(7.4)
a_nF
PEOE_VSA-3
PEOE_VSA-4
b_ar
LogD(5.5)
a_ICM
a_aro
vsurf_CW2
vsurf_R
LogD(6.5)
Q_VSA_FPPOS
opr_nring
pMDR1_J48-GA


MRP1 (CC)
x - misprediction
o - correct prediction
1.5
1.0
0.5
0.0
Q_VSA_POL
b_max1len
vsurf_D7
Q_VSA_FPPOS
b_rotR
rings
pMRP2_RfF


MRP2 (CC)
x - misprediction
o - correct prediction
1.5
1.0
0.5
0.0
Fi(B)
ast_violation_ext
reactive
b_rotR
Q_VSA_FHYD
vsurf_HB2
PEOE_VSA_FPNEG
vsurf_CW2

11.2. Appendix II: Supporting Information for Chapter 5

Figure A2.1-6. Applicability domain profiles for the CC model, where accuracy is the ratio of correct predictions over the total instances in the TE set. Each point in every graph is labelled with the amount of data being included within the current AD threshold.

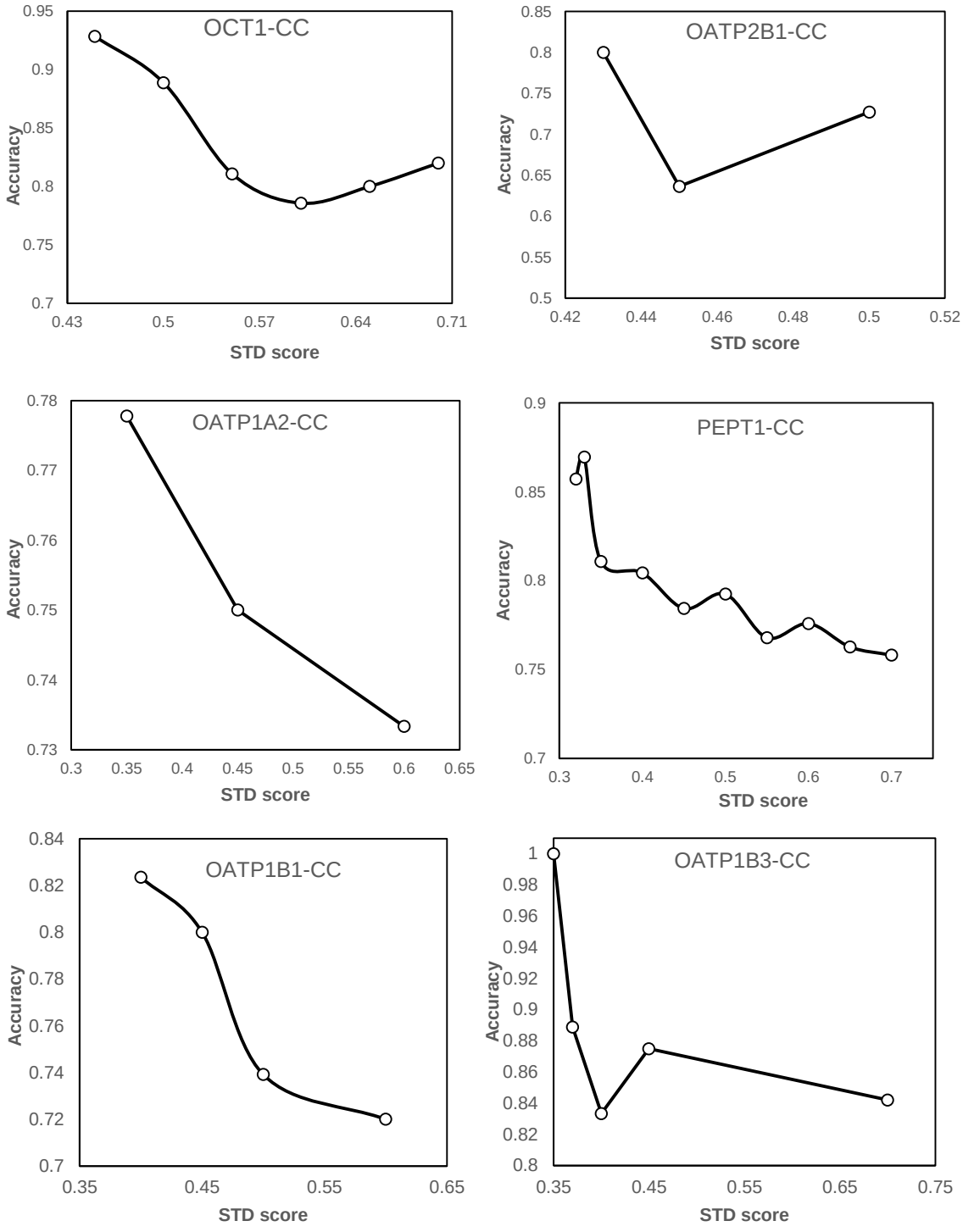


Table A2.1. Activity cliff (AC) rate in the different transporter datasets used in this work.

Model	%ACs among mispredictions of the BR model	%ACs among mispredictions of the CC model
OCT1	22% (**2** out of 9)	
OATP2B1	0 %	0 %
OATP1A2	0 %	0 %
PEPT1	43% (**9** out of 21)	46% (**7** out of 15)
OATP1B1	14% (**1** out of 7)	0 %
OATP1B3	0 %	0 %

Table A2.2. Chi-square test of pairwise correlation between observed transport data.

	1A2	**1B3**	**2B1**	**OCT1**	**PEPT1**
1B1	**0.026**	**< 0.001**	0.721	0.524	**0.008**
1A2		**0.014**	1.00	0.203	0.333
1B3			1.00	0.400	null
2B1				1.00	1.00
OCT1					null

Table A2.3. Chi-square test of correlation between observed and predicted transport data, where predicted data in used as output by the best multi-label model. O_S: observed substrate; O_NS: observed non-substrate; P_S: predicted substrate; P_NS: predicted non-substrate.

Observed \ Predicted	pOCT1	pOATP2B1	pOATP1A2	pPEPT1	pOATP1B1
OATP2B1	p = 0.876 P_S P_NS O_S 20 27 O_NS 27 37				
OATP1A2	p = 0.158 P_S P_NS O_S 34 21 O_NS 10 14	p = 0.948 P_S P_NS O_S 24 31 O_NS 11 13			
PEPT1	**p < 0,001** P_S P_NS O_S 218 28 O_NS 55 26	p = 0.110 P_S P_NS O_S 112 134 O_NS 28 53	p = 0.580 P_S P_NS O_S 206 40 O_NS 65 16		
OATP1B1	p = 0.649 P_S P_NS O_S 34 61 O_NS 11 26	p = 0.087 P_S P_NS O_S 27 68 O_NS 17 20	p = 0.667 P_S P_NS O_S 72 23 O_NS 26 11	p =0.678 P_S P_NS O_S 31 64 O_NS 10 27	
OATP1B3	p = 0.942 P_S P_NS O_S 18 40 O_NS 9 17	p = 0.564 P_S P_NS O_S 17 41 O_NS 10 16	p = 0.154 P_S P_NS O_S 44 14 O_NS 15 11	p = 0.102 P_S P_NS O_S 11 47 O_NS 10 16	**p < 0,0001*** P_S P_NS O_S 39 19 O_NS 3 23

Table A2.4. Descriptor importance for the BR model, measured in percentages of predicted and correctly predicted instances covered by each of the descriptors. For the sake of simplicity this table only shows up to the 10th most important feature, however some models used more features, as shown in Supporting Table A2.7. Their definitions are available in Supporting Table A2.6.

OCT1			OATP2B1			OATP1A2		
Descriptors	% N	% N correct	Descriptors	% N	% N correct	**Descriptors**	**% N**	**% N correct**
CASA-	100.0	96.5	PSA	100.0	93.0	vsurf_EDmin2	29.8	28.6
LogD7.4	88.4	84.8	vsurf_HB2	82.8	75.8	Nratio	25.5	25.0
PM3_dipole	82.0	78.5	PEOE_VSA_FPNEG	72.0	65.0	NumRings6	25.1	23.9
a_aro	39.8	38.4	PEOE_VSA_FHYD	56.7	51.2	Fu	23.4	22.2
vsurf_HB6	32.3	30.3	a_don	39.7	34.3	LogD5.5	20.2	19.6
lip_violation	22.8	21.4	AM1_E	33.6	28.1	SlogP_VSA1	20.1	19.4
vsurf_ID2	16.5	15.8				vsurf_DD13	18.9	18.0
FCASA+	4.5	3.8				a_don	16.0	15.3
Q_VSA_PNEG	4.1	4.1				FASA_H	11.2	10.6
						PEOE_VSA-1	10.8	10.3
PEPT1			**OATP1B1**			**OATP1B3**		
Descriptors	**% N**	**% N correct**	**Descriptors**	**% N**	**% N correct**	**Descriptors**	**% N**	**% N correct**
AM1_HF	93.3	83.1	FiA	34.4	34.4	vsurf_ID6	28.8	28.8
ast_violation_ext	48.9	42.8	PEOE_VSA_NEG	33.7	33.7	vsurf_ID5	24.7	24.6
SlogP_VSA6	44.5	39.9	vsurf_ID7	20.6	20.6	vsurf_ID1	21.1	21.1
Ro5	35.0	30.5	vsurf_EDmin2	18.5	18.5	ast_violation	18.7	18.6
Fu	19.2	17.2	Q_VSA_FPPOS	16.7	16.7	vsurf_ID2	18.6	18.5
FiA	18.9	17.3	SMR	16.4	16.3	NumRings6	17.4	17.4
a_nO	14.1	12.7	vdw_vol	16.2	16.2	vsurf_ID7	16.8	16.8
PSA	13.0	11.6	Q_VSA_FPOS	15.1	15.1	b_1rotN	16.1	16.1
a_acc	11.1	10.5	vsurf_CW2	14.9	14.9	AM1_Eele	14.3	14.3
a_hyd	11.1	10.2	vsurf_Wp6	14.8	14.8	Index of Refraction	14.1	14.1

Table A2.5. Descriptor importance for the CC model, measured in percentages of predicted and correctly predicted instances covered by each of the descriptors. For the sake of simplicity this table only shows up to the 10th most important feature, however some models used more features, as shown in Supporting Table A2.7. Their definitions are available in Supporting Table A2.6.

OCT1			OATP2B1			OATP1A2		
Descriptors	**% N**	**% N correct**	**Descriptors**	**% N**	**% N correct**	**Descriptors**	**% N**	**% N correct**
CASA-	100.0	96.5	PSA	100.0	96.9	vsurf_EDmin2	26.8	26.2
LogD7.4	88.4	84.8	vsurf_HB2	82.8	79.7	Fu	25.4	24.3
PM3_dipole	82.0	78.5	PEOE_VSA_FPNEG	72.0	69.0	NumRings6	25.1	24.3
a_aro	39.8	38.4	PEOE_VSA_FHYD	56.7	55.1	**pOCT1**	24.3	23.4
vsurf_HB6	32.3	30.3	a_don	39.7	38.2	a_don	24.2	23.3
lip_violation	22.8	21.4	**pOCT1**	33.6	32.0	SlogP_VSA1	16.1	16.0
vsurf_ID2	16.5	15.8	PM3_dipole	21.3	21.3	FASA_H	13.2	12.9
FCASA+	4.5	3.8				Nratio	12.0	11.5
Q_VSA_PNEG	4.1	4.1				PEOE_VSA-1	10.9	10.7
						LogP	9.2	9.1
PEPT1			**OATP1B1**			**OATP1B3**		
Descriptors	**% N**	**% N correct**	**Descriptors**	**% N**	**% N correct**	**Descriptors**	**% N**	**% N correct**
AM1_HF	67.0	62.2	FiA	39.0	39.0	**pOATP1B1**	26.5	26.5
pOCT1	57.1	53.2	PEOE_VSA_NEG	28.1	28.1	vsurf_ID6	25.6	25.6
FiA	42.9	39.9	vsurf_ID7	18.3	18.3	vsurf_ID1	18.6	18.6
pOATP2B1	41.8	38.8	vdw_vol	17.8	17.8	vsurf_ID2	18.3	18.2
Ro5	26.2	24.2	Q_VSA_FPOS	17.5	17.5	vsurf_ID5	17.6	17.6
ast_violation_ext	19.2	17.8	Q_VSA_FPPOS	17.1	17.1	NumRings6	15.6	15.6
pOATP1A2	14.9	13.8	SMR	16.4	16.4	vsurf_R	13.6	13.6
LogD7.4	13.7	13.0	vsurf_EDmin2	16.2	16.2	FRB[#]	12.4	12.4
SlogP_VSA6	12.3	11.5	glob	16.1	16.1	ast_violation	12.0	12.0
FiAB	10.5	9.7	vdw_area	14.5	14.5	b_1rotN[#]	11.6	11.6

Both count single bonds, however using different parameters. The full definitions of the descriptors are available in Table A2.6.

Table A2.6. Definitions of all molecular descriptors present in the BR and CC models.

Feature	Models where the feature is present	Definition
a_acc	PEPT1-BR PEPT1-CC	Number of hydrogen bond acceptor atoms (not counting acidic atoms but counting atoms that are both hydrogen bond donors and acceptors such as -OH).
a_aro	OCT1(-BR/CC) OATP1B1-BR OATP1B1-CC	Aromatic atoms
a_don	OATP1A2-BR OATP2B1-BR OATP1A2-CC OATP2B1-CC	Number of hydrogen bond donor atoms (not counting basic atoms but counting atoms that are both hydrogen bond donors and acceptors such as -OH)
a_hyd	OATP1B3-BR PEPT1-BR OATP1B3-CC PEPT1-CC	Number of hydrophobic atoms.
a_nCl	OATP1A2-BR OATP1A2-CC	Number of chlorine atoms
a_nO	OATP1B1-BR PEPT1-BR OATP1B1-CC PEPT1-CC	Number of oxygen atoms

Table A2.6. (Cont.)

Feature	Models where the feature is present	Definition
AM1_E	OATP1B3-BR OATP2B1-BR OATP1B3-CC	The total Self-consistent field energy (kcal/mol) calculated using the AM1 Hamiltonian.
AM1_Eele	OATP1B3-BR OATP1B3-CC	The electronic energy (kcal/mol) calculated using the AM1 Hamiltonian.
AM1_HF	PEPT1-BR PEPT1-CC	The heat of formation (kcal/mol) calculated using the AM1 Hamiltonian.
ASA_P	OATP1A2-BR OATP1A2-CC	Total polar surface area
ast_violation	OATP1B3-BR OATP1B3-CC	Astex violations, otherwise known as the rule of 3
ast_violation_ext	PEPT1-BR PEPT1-CC	Astex violations (extended), otherwise known as the rule of 3
b_1rotN	OATP1B3-BR OATP1B3-CC	Number of rotatable single bonds. Conjugated single bonds are not included (e.g. ester and peptide bonds).
b_heavy	OATP1B3-BR OATP1B3-CC	Number of heavy-heavy bonds
CASA-	OCT1(-BR/CC)	Negative charge weighted surface area, ASA- multiplied by the maximum negative partial charge.
FASA_H	OATP1A2-BR OATP1A2-CC	Fractional ASA_H calculated as ASA_H / ASA.
FASA+	OATP1B1-BR OATP1B1-CC	Fractional positive accessible surface area
FCASA+	OCT1(-BR/CC)	Fractional CASA+ calculated as CASA+ / ASA. The negative equivalent of FCASA- definition.
FiA	OATP1A2-BR OATP1B1-BR PEPT1-BR OATP1A2-CC OATP1B1-CC PEPT1-CC	Fraction of ionised acidic species at pH 7.4
FiAB	PEPT1-BR PEPT1-CC	Fraction of ionized zwitterionic species, where both acidic and basic moieties are ionized.
FiB	PEPT1-BR PEPT1-CC	Fraction of ionised basic species at pH 7.4
FRB	OATP1B3-BR PEPT1-BR OATP1B3-CC PEPT1-CC	Freely Rotatable Bonds
Fu	OATP1A2-BR PEPT1-BR OATP1A2-CC PEPT1-CC	Unionized fraction at pH 7.4
glob	OATP1B1-BR OATP1B1-CC	Globularity. The ratio of surface to the surface of a same-volume sphere.
HAcceptors	OATP1B1-BR OATP1B1-CC	Hydrogen acceptors (N, O, and F atoms with free lone pairs of electrons)
IndexofRefraction	OATP1B3-BR OATP1B3-CC	Ratio of the speed of light in a vacuum to the speed of light in a medium under consideration.
lip_violation	OCT1(-BR/CC)	The number of violations of Lipinski's Rule of Five.
LogD10	OATP1B3-BR PEPT1-BR OATP1B3-CC PEPT1-CC	Log(octanol/water at pH 10)
LogD2	PEPT1-BR PEPT1-CC	Log(octanol/water at pH 2)
LogD5.5	OATP1A2-BR PEPT1-BR OATP1A2-CC PEPT1-CC	Log(octanol/water at pH 5.5)

Table A2.6. (Cont.)

Feature	Models where the feature is present	Definition
LogD6.5	PEPT1-CC	Log(octanol/water at pH 6.5)
LogD7.4	OCT1(-BR/CC) OATP1B1-BR PEPT1-BR OATP1B1-CC PEPT1-CC	Log(octanol/water at pH 7.4)
LogP	OATP1A2-BR PEPT1-BR OATP1A2-CC PEPT1-CC	Log(octanol/water)
NOratio	PEPT1-BR PEPT1-CC	Ratio of Nitrogen + Oxygen atoms (?)
Nratio	OATP1A2-BR OATP1A2-CC	Ratio of Nitrogen atoms
NumRings6	OATP1B3-BR OATP1A2-BR OATP1B3-CC OATP1A2-CC	Number of 6-membered rings
PEOE_VSA-4	OATP1B1-BR OATP1B1-CC	Sum of VDW volume of atoms where the atomic charge is in the range [-0.25,-0.20).
PEOE_VSA_FHYD	OATP2B1-BR OATP2B1-CC	Fractional hydrophobic van der Waals surface area. This is the sum of VSA of atoms with partial charge less than or equal to 0.2 divided by the total surface area.
PEOE_VSA_FPNEG	OATP2B1-BR OATP2B1-CC	Fractional negative polar van der Waals surface area. The sum of VSA for atoms with atomic charge < -0.2, divided by the total surface area.
PEOE_VSA_NEG	OATP1B1-BR OATP1B1-CC	Total negative van der Waals surface area. The sum VDW contribution from atoms with negative atomic charge.
PEOE_VSA_PPOS	OATP1B1-BR OATP1B1-CC	Total polar positive vdw surface area
PEOE_VSA+4	OATP1A2-BR OATP1A2-CC	Sum of VDW volume of atoms where the atomic charge is in the range [0.20,0.25)
PEOE_VSA+5	OATP1B1-BR OATP1B1-CC	Sum of VDW volume of atoms where the atomic charge is in the range [0.25,0.30).
PEOE_VSA+6	OATP1B1-BR OATP1B1-CC	Sum of VDW volume of atoms where the atomic charge is greater than 0.3.
PEOE_VSA-1	OATP1A2-BR OATP1A2-CC	Sum of VDW volume of atoms where the atomic charge is in the range [-0.10,-0.05).
PM3_dipole	OCT1(-BR/CC) OATP2B1-CC	The dipole moment calculated using the PM3 Hamiltonian.
PM3_LUMO	OATP1B1-BR OATP1B1-CC	The energy (eV) of the Lowest Unoccupied Molecular Orbital calculated using the PM3 Hamiltonian.
pOATP1A2	OATP1B3-CC PEPT1-CC	Predicted OATP1A2 binding, output by the OATP1A2 single-label model trained within the best multi-label model.
pOATP1B1	OATP1B3-CC	Predicted OATP1B1 binding, output by the OATP1B1 single-label model trained within the best multi-label model.
pOATP2B1	OATP1B3-CC OATP1A2-CC PEPT1-CC	Predicted OATP2B1 binding, output by the OATP2B1 single-label model trained within the best multi-label model.
pOCT1	OATP1B3-CC OATP1A2-CC OATP2B1-CC PEPT1-CC	Predicted OCT1 binding, output by the OCT1 single-label model trained in isolation (as it is used as the first label of the CC models).
pPEPT1	OATP1B3-CC	Predicted PEPT1 binding, output by the PEPT1 single-label model trained within the best multi-label model.
PSA	OATP2B1-BR PEPT1-BR OATP2B1-CC PEPT1-CC	Polar Surface area. Measure of how much exposed polar area a molecule has.

Table A2.6. (Cont.)

Feature	Models where the feature is present	Definition
Q_VSA_FHYD	OATP1A2-BR OATP1A2-CC	Fractional hydrophobic vdw surface area. This is the sum of VSA of atoms with partial charge less than or equal to 0.2 divided by the total surface area.
Q_VSA_FPOS	OATP1B1-BR OATP1B1-CC	Fractional positive van der Waals surface area. This is the sum of the vi such that qi is non-negative divided by the total surface area. The vi are calculated using a connection table approximation.
Q_VSA_FPPOS	OATP1B1-BR OATP1B1-CC	Fraction of positive polar VSA (PM6-derived) defined as the van der Waals area of atoms with atomic charge > 0.2, divided by the total surface area.
Q_VSA_PNEG	OCT1(-BR/CC)	Total negative polar van der Waals surface area. This is the sum of the vi such that qi is less than -0.2. The vi are calculated using a connection table approximation.
Ro5	PEPT1-BR PEPT1-CC	Rule of Five (H-bond donors <= 5; H-bond acceptors <= 10; MW < 500; logP < 5).
SlogP_VSA1	OATP1A2-BR OATP1A2-CC	VSA of atoms which contributing logP is (-0.4,-0.2].*
SlogP_VSA6	PEPT1-BR PEPT1-CC	VSA of atoms which contributing logP is (0.20,0.25].*
SMR	OATP1B1-BR OATP1B1-CC	Molecular refractivity (including implicit hydrogens). Measures the volume occupied per mol of substance, and carries information about volume and polarizability.
SMR_VSA4	OATP1B1-BR OATP1B1-CC	Total VDW surface area of atoms with molecular refractivity ranging (0.39,0.44]*
vdw_area	OATP1B1-BR OATP1B1-CC	Area of van der Waals surface (A^2) calculated using a connection table approximation.
vdw_vol	OATP1B1-BR OATP1B1-CC	van der Waals volume (A^3) calculated using a connection table approximation.
vsa_hyd	OATP1B3-BR OATP1B3-CC	VDW hydrophobe surface area (A^2)
vsa_other	OATP1B1-BR OATP1B1-CC	VDW other surface area (A^2)
vsurf_CW2	OATP1B1-BR OATP1B1-CC	Capacity factor, calculated as the ratio of the hydrophilic surface to the total molecular surface, at -0.5 kcal/mol.
vsurf_CW6	OATP1A2-BR OATP1A2-CC	Capacity factor, calculated as the ratio of the hydrophilic surface to the total molecular surface, at -4.0 kcal/mol.
vsurf_DD13	OATP1A2-BR OATP1A2-CC	Contact distances of vsurf_DDmin.
vsurf_EDmin2	OATP1A2-BR OATP1B1-BR OATP1A2-CC OATP1B1-CC	Hydrophobic Local Interaction Energy Minima
vsurf_HB2	OATP2B1-BR OATP2B1-CC	Hydrogen-bond donor capacity. Defined as: Volume $_{\text{hydrophilic interactions}}$ − Volume $_{\text{O probe interactions}}$ (at -0.5 kcal/mol).
vsurf_HB6	OCT1(-BR/CC)	Hydrogen-bond donor capacity. Defined as: Volume $_{\text{hydrophilic interactions}}$ − Volume $_{\text{O probe interactions}}$ (at -4.0 kcal/mol).
vsurf_HB8	OATP1B1-BR OATP1B1-CC	Hydrogen -bond donor capacity. Defined as: Volume $_{\text{hydrophilic interactions}}$ − Volume $_{\text{O probe interactions}}$ (at -6.0 kcal/mol).
vsurf_HL1	OATP1B1-BR OATP1B1-CC	First hydrophilic-lipophilic balance. Ratio of the volume of hydrophobic regions to the volume of hydrophobic regions.
vsurf_ID1	OATP1B3-BR OATP1B3-CC	Hydrophobic integy moment at an energy level of -0.2 kcal/mol.
vsurf_ID2	OCT1(-BR/CC) OATP1B3-BR OATP1B3-CC	Hydrophobic integy moment at an energy level of -0.4 kcal/mol.
vsurf_ID3	OATP1B3-BR OATP1B3-CC	Hydrophobic integy moment at an energy level of -0.6 kcal/mol.

Table A2.6. (Cont.)

Feature	Models where the feature is present	Definition
vsurf_ID4	OATP1B3-BR OATP1B3-CC	Hydrophobic integy moment at an energy level of -0.8 kcal/mol.
vsurf_ID5	OATP1B3-BR OATP1B3-CC	Hydrophobic integy moment at an energy level of -1.0 kcal/mol.
vsurf_ID6	OATP1B3-BR OATP1B3-CC	Hydrophobic integy moment at an energy level of -1.2 kcal/mol.
vsurf_ID7	OATP1B3-BR OATP1B1-BR OATP1B3-CC OATP1B1-CC	Hydrophobic integy moment at an energy level of -1.4 kcal/mol.
vsurf_R	OATP1B3-BR OATP1B3-CC	Surface rugosity, defined as the ratio of volume to surface
vsurf_Wp3	OATP1B1-BR OATP1B1-CC	Polar volume. Volume of the interactions with an O probe at -1 kcal/mol
vsurf_Wp6	OATP1B1-BR OATP1B1-CC	Polar volume. Volume of the interactions with an O probe at -4 kcal/mol
Weight	OATP1B1-BR OATP1B1-CC	Molecular Weight (g/mol).
weinerPol	OATP1B3-BR OATP1B3-CC	Weiner polarity number

*Atom contribution values listed on Wildman and Crippen(Wildman and Crippen, 1999).

Table A2.7. Full set of molecular descriptors that compose each multi-label model.

BR model					
OCT1	OATP1B3	OATP1A2	OATP2B1	OATP1B1	PEPT1
CASA- LogD7.4 PM3_dipole a_aro vsurf_HB6 lip_violation vsurf_ID2 Q_VSA_PNEG FCASA+	vsurf_ID6 vsurf_ID5 vsurf_ID1 ast_violation vsurf_ID2 NumRings6 vsurf_ID7 b_1rotN AM1_Eele IndexofRefraction vsurf_R FRB weinerPol LogD10 vsurf_ID3 AM1_E a_hyd vsurf_ID4 b_heavy vsa_hyd	vsurf_EDmin2 Nratio NumRings6 Fu LogD5.5 SlogP_VSA1 vsurf_DD13 a_don FASA_H PEOE_VSA-1 Q_VSA_FHYD PEOE_VSA+4 FiA ASA_P LogP a_nCl vsurf_CW6	PSA vsurf_HB2 PEOE_VSA_FPNEG PEOE_VSA_FHYD a_don AM1_E	FiA PEOE_VSA_NEG vsurf_ID7 vsurf_EDmin2 Q_VSA_FPPOS SMR vdw_vol Q_VSA_FPOS vsurf_CW2 vsurf_Wp6 vdw_area Weight glob LogD7.4 vsurf_Wp3 FASA+ a_nO PEOE_VSA_PPOS vsa_other vsurf_HL1 PEOE_VSA+6 vsurf_HB8 PM3_LUMO SMR_VSA4 HAcceptors PEOE_VSA.4 PEOE_VSA+5 a_aro	AM1_HF ast_violation_ext SlogP_VSA6 Ro5 FiA Fu a_nO PSA a_acc a_hyd FRB LogD2 LogP FiB LogD5.5 FiAB NOratio LogD7.4 LogD10
CC model					
OCT1	OATP1B3	OATP1A2	OATP2B1	OATP1B1	PEPT1
(same as above)	pOATP1B1 vsurf_ID6 vsurf_ID1 vsurf_ID2 vsurf_ID5 NumRings6 vsurf_R FRB ast_violation b_1rotN vsurf_ID7 vsurf_ID3 IndexofRefraction AM1_Eele vsurf_ID4 a_hyd weinerPol LogD10 pOATP1A2 AM1_E vsa_hyd b_heavy pOCT1 pOATP2B1 pPEPT1	vsurf_EDmin2 Fu NumRings6 pOCT1 a_don SlogP_VSA1 FASA_H Nratio PEOE_VSA-1 LogP ASA_P FiA LogD5.5 PEOE_VSA+4 vsurf_DD13 a_nCl vsurf_CW6 pOATP2B1 Q_VSA_FHYD	PSA vsurf_HB2 PEOE_VSA_FPNEG PEOE_VSA_FHYD a_don pOCT1 PM3_dipole	FiA PEOE_VSA_NEG vsurf_ID7 vdw_vol Q_VSA_FPOS Q_VSA_FPPOS SMR vsurf_EDmin2 glob vdw_area Weight vsurf_Wp6 LogD7.4 vsurf_CW2 PEOE_VSA+6 FASA+ vsurf_Wp3 vsurf_HL1 SMR_VSA4 vsa_other vsurf_HB8 a_nO PEOE_VSA_PPOS PM3_LUMO PEOE_VSA.4 a_aro HAcceptors PEOE_VSA+5 pOATP1A2 pPEPT1 pOATP2B1 pOCT1	AM1_HF pOCT1 FiA pOATP2B1 Ro5 ast_violation_ext pOATP1A2 LogD7.4 SlogP_VSA6 FiAB PSA a_acc a_hyd LogD5.5 LogD10 a_nO LogD2 FRB NOratio LogD6.5 FiB LogP Fu

Table A2.8. Expression levels of different SLC transporters across a wide range of tissues. These expression levels for all transporters except OATP1A2 are obtained from western blot quantification reported in the protein atlas platform(Uhlén et al., 2015) (http://www.proteinatlas.org/) and reported in low (L), medium (M) or high (H) quantified amount of protein. For these, empty cells mean that the presence of the transporter in the tissue was tested, and yielded non-detectable amount of protein. On the other hand, the protein expression for OATP1A2 is not available from protein atlas, and instead has been gathered from the western blot analysis reported in the literature(Franke et al., 2009, Lee et al., 2005). In the case of OATP1A2, measurements are annotated with "Y" for observed transporter in the tissue, "n.m" for no reported information found, and empty cells for tested but non detectable expression.

	OATP1B1	OATP2B1	OATP1B3	OCT1	PEPT1	OATP1A2
brain		M		L		Y
lateral ventricle						n.m.
thyroid gland		L		L		n.m.
parathyroid gland		L		L		n.m.
adrenal gland		M		M		n.m.
appendix						n.m.
bone marrow		L		L		n.m.
tonsil						n.m.
heart muscle		M		M		n.m.
skeletal muscle		L		M		n.m.
nasopharynx				L		n.m.
Lung		L		L		n.m.
liver	M	L	H	M		
gallbladder				M	M	n.m.
pancreas		M		L		n.m.
esophagus				M		n.m.
stomach		M		M		n.m.
duodenum				M	M	Y
small intestine				M	M	n.m.
colon				M		n.m.
rectum				M		n.m.
kidney				M		Y
urinary bladder				L		n.m.
testis		H		L		n.m.
seminal vesicle						n.m.
breast		L				Y
cervix, uterine				H		n.m.
endometrium						n.m.
fallopian tube				L		n.m.
ovary		L		L		n.m.
placenta				M		n.m.
adipose tissue						n.m.
soft tissue						n.m.
skin				M		n.m.

11.3. Appendix III: Supporting Information for Chapter 6

Feature selection from **ALL FEATURES**

Table A3.1. Models built from all available descriptors submitted to a prior run of GS pre-processing feature selection.

	Greedy Search FS			
	Random Forest		Boosted Trees	
	ePL (1)	pPL (1a)	ePL (12)	pPL (12a)
MAE	0,3385	0,3385	0,441	0,441
	FiA (86.4) vsurf_CW5 (79.1) vsurf_CP (73.2) FiB (71.7) Q_VSA_FPNEG (66.1) glob (64.9) AM1_dipole (61.1) vsurf_CW8 (58.8) vsurf_DD13 (45.1) PEOE_VSA-2 (33.6) vsurf_DW13 (32.6)	FiA (86.4) vsurf_CW5 (79.1) vsurf_CP (73.2) FiB (71.7) Q_VSA_FPNEG (66.1) glob (64.9) AM1_dipole (61.1) vsurf_CW8 (58.8) vsurf_DD13 (45.1) PEOE_VSA-2 (33.6) vsurf_DW13 (32.6)	FiA (96.5) FiB (78.2) vsurf_CW5 (76.6) glob (76.3) AM1_dipole (70) vsurf_CW8 (68.9) Q_VSA_FPNEG (61.9) vsurf_CP (61.5) vsurf_DD13 (47.5) vsurf_DW13 (28.9) PEOE_VSA-2 (21.6)	FiA (96.5) FiB (78.2) vsurf_CW5 (76.6) glob (76.3) AM1_dipole (70) vsurf_CW8 (68.9) Q_VSA_FPNEG (61.9) vsurf_CP (61.5) vsurf_DD13 (47.5) vsurf_DW13 (28.9) PEOE_VSA-2 (21.6)

Table A3.2. Models built from all available descriptors submitted to a prior run of GA pre-processing feature selection.

	Genetic Search FS			
	Random Forest		Boosted Trees	
	ePL (5)	pPL (5a)	ePL (16)	pPL (16a)
MAE	0,317	**0,308**	0.322	0.304
	FiA (47.8) LogD(10) (44.7) ASA_P (35) Hetero_ratio (33.9) AM1_HOMO (31.8) FASA_H (31.4) PEOE_VSA_FHYD (31) vsurf_HB2 (29.5) Q_VSA_FHYD (26.4) vsurf_CP (26.3) PEOE_VSA-0 (26.2) LogD(6_5) (26) **pPEPT1 (25.8)** SMR_VSA0 (25.4) LogP (25.4) vsurf_HB1 (24.7) vsurf_CW2 (23.9) SMR_VSA2 (23.8) vsurf_HB3 (23.8) SlogP_VSA5 (23.7) Surface_Tension (23.1) PEOE_RPC- (22.4) C_ratio (22.1) vsurf_EWmin3 (21.8) vsurf_IW3 (20.6) Q_VSA_FPOS (20) balabanJ (19.8) vsurf_EDmin3 (18.1) vsurf_DD13 (16.7) PEOE_VSA+5 (15.9) **pMRP2 (15)** a_acc (12.9) PEOE_VSA-2 (12.6) vsurf_DD23 (11.6) vsurf_DD12 (10.4) Num_Rings_5 (8.2) chiral_u (6.7) Halogen_ratio (6.5) PEOE_VSA+6 (6.3) **pOATP2B1 (5.3)** **pOATP1B1 (4.9)** **ePL (0.3)** density (0)	LogD(10) (51.8) FiA (50.6) vsurf_HL2 (36.1) ASA_P (33.3) vsurf_HL1 (31.9) SMR_VSA0 (31.8) AM1_HOMO (30.4) vsurf_HB1 (29) **pPEPT1 (26.9)** PEOE_VSA-0 (26.5) C_ratio (26.2) vsurf_HB3 (25.2) vsurf_HB2 (25.1) LogP (25) vsurf_W6 (24.8) NO_ratio (24.2) PEOE_RPC+ (24.1) SMR_VSA2 (23.6) vsurf_CW3 (23) Surface_Tension (22.6) vsurf_ID3 (21.3) vsurf_IW5 (20.7) MNDO_dipole (20.6) **pMRP2 (20.4)** vsurf_EDmin2 (20.2) PM3_dipole (18.9) Density (17.8) vsurf_DD13 (17.3) **pPL (15.9)** **pOCT1 (13.6)** PEOE_VSA-2 (13.4) vsurf_DW13 (12.6) a_nN (12.2) vsurf_DW23 (12.2) **pMRP1 (11.5)** vsurf_DD12 (11.1) Halogen_ratio (11.1) vsurf_DW12 (10.9) a_nS (10.6) FiAB (7) ast_violation (6.9) PEOE_VSA+6 (5) Num_Rings_4 (3.3) **pOATP1B1 (3.1)** Num_Rings_3 (1.7)	FiA (66.3) LogD(10) (56.4) Hetero_ratio (40.3) AM1_HOMO (40.2) FASA_H (39.2) PEOE_VSA_FHYD (37.8) ASA_P (35.7) vsurf_CP (31.3) SlogP_VSA5 (30.9) LogP (30.6) SMR_VSA0 (29.7) PEOE_VSA-0 (26.7) vsurf_HB2 (26.5) **pPEPT1 (25.4)** PEOE_RPC- (25.2) LogD(6_5) (25.1) Q_VSA_FHYD (23.8) Surface_Tension (23.5) SMR_VSA2 (22.1) vsurf_EDmin3 (22.1) Q_VSA_FPOS (21.8) density (21.2) vsurf_HB1 (21.1) vsurf_IW3 (21.1) balabanJ (21) vsurf_CW2 (20.5) vsurf_EWmin3 (19.9) PEOE_VSA+5 (18.9) C_ratio (18.7) vsurf_HB3 (18.3) vsurf_DD13 (17.3) **pMRP2 (15.3)** a_acc (12.8) PEOE_VSA-2 (12.3) vsurf_DD23 (10.6) vsurf_DD12 (8.8) Num_Rings_5 (7.8) chiral_u (5.9) Halogen_ratio (5.7) **pOATP2B1 (5)** PEOE_VSA+6 (4.9) **pOATP1B1 (4.5)** **ePL (0.2)**	FiA (60) LogD(10) (58.9) ASA_P (42.9) AM1_HOMO (40.6) vsurf_HL2 (40) SMR_VSA0 (37.3) vsurf_HL1 (35.8) PEOE_VSA-0 (32.2) NO_ratio (31.6) LogP (30.9) vsurf_W6 (29) vsurf_HB1 (28.8) C_ratio (28.5) SMR_VSA2 (28.1) vsurf_HB2 (27.6) PEOE_RPC+ (27.3) vsurf_HB3 (25.9) **pPEPT1 (23)** vsurf_ID3 (22.7) **pMRP2 (22.2)** vsurf_EDmin2 (22) vsurf_IW5 (20.4) MNDO_dipole (20.2) Surface_Tension (20.1) vsurf_CW3 (19.2) PM3_dipole (18.6) **pPL (16.5)** **pOCT1 (15.3)** vsurf_DD13 (15.1) Density (14.1) PEOE_VSA-2 (13.5) a_nN (13.5) vsurf_DW23 (13.1) a_nS (12.2) vsurf_DW13 (12.2) **pMRP1 (12.1)** vsurf_DW12 (11.8) vsurf_DD12 (9.7) FiAB (9.4) Halogen_ratio (8.2) ast_violation (6.3) PEOE_VSA+6 (4.4) **pOATP1B1 (4)** Num_Rings_4 (3.5) Num_Rings_3 (2.2)

Feature selection from **Physiological Descriptors**

Table A3.3. Models built from physiological descriptors (exclusively), selected in a prior run of GS pre-processing feature selection.

	Greedy Search FS			
	Random Forest		Boosted Trees	
	ePL (3)	pPL (3a)	ePL (14)	pPL (14a)
MAE	0,469	**0,424**	0,599	**0,588**
	pPEPT1 (97.2) pBCRP1 (93.7) pMRP2 (93.4) pMRP1 (89.6) pOATP1B1 (63.7) **ePL (30.3)**	pPEPT1 (95.6) **pPL (93.1)** pMRP2 (89.7) pBCRP1 (86.6) pMRP1 (84.4) pOATP1B1 (42.8)	pMRP2 (99.3) pPEPT1 (97.6) pBCRP1 (95.5) pMRP1 (82.7) pOATP1B1 (61.9) **ePL (38.7)**	pMRP2 (99.7) **pPL (95.8)** pPEPT1 (94.1) pBCRP1 (81.9) pOATP1B1 (75.1) pMRP1 (71.7)

Table A3.4. Models built from physiological descriptors (exclusively), selected in a prior run of GA pre-processing feature selection.

	Genetic Search FS			
	Random Forest		Boosted Trees	
	ePL (7)	pPL (7a)	ePL (18)	pPL (18a)
MAE	0,469	**0,424**	0,599	**0,588**
	pPEPT1 (97.2) pBCRP1 (93.7) pMRP2 (93.4) pMRP1 (89.6) pOATP1B1 (63.7) **ePL (30.3)**	pPEPT1 (95.6) **pPL (93.1)** pMRP2 (89.7) pBCRP1 (86.6) pMRP1 (84.4) pOATP1B1 (42.8)	pMRP2 (99.3) pPEPT1 (97.6) pBCRP1 (95.5) pMRP1 (82.7) pOATP1B1 (61.9) **ePL (38.7)**	pMRP2 (99.7) **pPL (95.8)** pPEPT1 (94.1) pBCRP1 (81.9) pOATP1B1 (75.1) pMRP1 (71.7)

Feature selection from **Physiological Descriptors** + Feature selection from **Molecular Descriptors**

Table A3.5. Models based on two parallel runs of GS pre-processing feature selection done on physiological descriptors and molecular descriptors separately. Both sets of selected descriptors were merged and used for training of the QSAR models.

	Greedy Search FS			
	Random Forest		Boosted Trees	
	ePL (4)	pPL (4a)	ePL (15)	pPL (15a)
MAE	0,3357	**0,3273**	**0,360**	0,370
	FiA (77.6) vsurf_CW5 (68.5) FiB (65.2) vsurf_CP (63.9) pPEPT1 (58.3) Q_VSA_FPNEG (54.3) glob (53) AM1_dipole (50.3) vsurf_CW8 (46.9) pBCRP1 (46.6) pMRP2 (37.1) pMRP1 (34.4) vsurf_DD13 (34.3) vsurf_DW13 (28.4) PEOE_VSA-2 (24.7) pOATP1B1 (10.9) **ePL (0.9)**	FiA (78) vsurf_CW5 (67.2) FiB (63.9) vsurf_CP (63.1) pPEPT1 (53) glob (51.4) Q_VSA_FPNEG (51.3) vsurf_CW8 (47.6) AM1_dipole (45.5) **pPL (43.5)** pBCRP1 (41.6) pMRP2 (40) pMRP1 (32) vsurf_DD13 (31.1) vsurf_DW13 (27) PEOE_VSA-2 (23.2) pOATP1B1 (9.1)	FiA (91.7) FiB (78.1) vsurf_CW5 (73.5) glob (65.3) vsurf_CP (59) Q_VSA_FPNEG (56.8) vsurf_CW8 (53.2) pPEPT1 (52.9) pBCRP1 (50.2) AM1_dipole (47.6) pMRP2 (38.2) pMRP1 (35.7) vsurf_DD13 (34) vsurf_DW13 (23.2) PEOE_VSA-2 (22.4) pOATP1B1 (5.6) **ePL (0.8)**	FiA (92) FiB (75.7) vsurf_CW5 (71.9) glob (61) vsurf_CP (59.7) Q_VSA_FPNEG (59.6) vsurf_CW8 (54.3) pPEPT1 (45.7) AM1_dipole (45.1) pBCRP1 (44.1) **pPL (42)** pMRP2 (34.1) pMRP1 (33.7) vsurf_DD13 (31) vsurf_DW13 (22.4) PEOE_VSA-2 (21.1) pOATP1B1 (4.9)

Table A3.6. Models based on two parallel runs of pre-processing GA feature selection done on physiological descriptors and molecular descriptors separately. Both sets of selected descriptors were merged and used for training of the QSAR models.

	Genetic Search FS			
	Random Forest		Boosted Trees	
	ePL (8)	pPL (8a)	ePL (19)	pPL (19a)
MAE	0,307	**0,306**	0,321	**0,316**
	FiA (45.2) LogD(10) (40.3) ASA_P (33.6) LogD(7_4) (31.8) FiB (31.8) Hetero_ratio (28.8) vsurf_HL1 (27.3) SMR_VSA0 (26.4) **pPEPT1 (25.3)** vsurf_HB2 (25) vsurf_CW6 (24.4) density (24.3) PEOE_VSA-0 (23.9) vsurf_HB1 (23.8) MNDO_HOMO (23.8) LogP (23.6) b_1rotR (22.7) vsurf_HB3 (21.7) PEOE_PC- (21.6) SlogP_VSA5 (21.4) VAdjEq (21) LogD(5_5) (20.5) Surface_Tension (19.3) vsurf_ID4 (19.1) vsa_acc (18.7) SlogP_VSA1 (18.2) dens (17.9) **pBCRP1 (17.7)** vsurf_EWmin3 (17.6) MNDO_LUMO (17.3) vsurf_ID5 (17) vsurf_W7 (16.9) vsurf_IW8 (16.1) vsurf_ID6 (16.1) **pMRP2 (15.7)** AM1_dipole (15.4) **pMRP1 (12.6)** vsurf_DW23 (11.1) PEOE_VSA-2 (10.5) vsurf_DD23 (10.4) Num_Rings_6 (8.8) Num_Rings_5 (5.5) **pOATP1B1 (4.1)** Num_Rings_4 (3.7) reactive (3.3) Num_Rings_3 (1.5) **ePL (0.2)**	FiA (47.1) LogD(10) (38.5) LogD(7_4) (31.9) FiB (30.8) ASA_P (30.6) Hetero_ratio (29.6) vsurf_HL1 (29) vsurf_CW6 (25.3) vsurf_HB1 (24.9) **pPEPT1 (24.5)** vsurf_HB2 (24.2) density (24) SMR_VSA0 (23.4) MNDO_HOMO (23.4) vsurf_HB3 (22.6) PEOE_VSA-0 (22.6) LogP (22.5) b_1rotR (21.1) SlogP_VSA5 (20.6) VAdjEq (20.2) PEOE_PC- (19.9) vsa_acc (19.9) LogD(5_5) (19.3) vsurf_ID4 (18.9) Surface_Tension (18.4) SlogP_VSA1 (18.4) **pBCRP1 (17.7)** MNDO_LUMO (17.2) **pPL (17.1)** vsurf_EWmin3 (17) vsurf_ID5 (16.9) dens (16.5) vsurf_W7 (16.3) vsurf_IW8 (15.5) **pMRP2 (15.3)** AM1_dipole (15.3) vsurf_ID6 (14.9) **pMRP1 (12.5)** vsurf_DD23 (11.4) vsurf_DW23 (10.9) PEOE_VSA-2 (9.6) Num_Rings_6 (9.5) Num_Rings_5 (6.1) **pOATP1B1 (3.8)** reactive (3.6) Num_Rings_4 (3.3) Num_Rings_3 (1.3)	FiA (54.9) LogD(10) (52.6) Hetero_ratio (43) FiB (40.6) ASA_P (35) vsurf_CW6 (31.9) LogD(7_4) (31.2) vsurf_HL1 (29.3) b_1rotR (27.3) vsurf_HB2 (27.2) MNDO_HOMO (27.1) SMR_VSA0 (27.1) SlogP_VSA5 (26.8) **pPEPT1 (25.5)** PEOE_VSA-0 (25.3) LogP (24.8) vsurf_HB1 (23.9) density (23.7) VAdjEq (22.6) MNDO_LUMO (22) vsurf_ID4 (20.7) PEOE_PC- (20.6) LogD(5_5) (20.6) vsurf_HB3 (20.1) **pBCRP1 (20.1)** SlogP_VSA1 (19.9) vsurf_EWmin3 (18.5) Surface_Tension (18.4) vsa_acc (17.7) vsurf_ID5 (17.5) dens (16.9) vsurf_IW8 (16.4) vsurf_W7 (16.2) **pMRP2 (15.5)** AM1_dipole (15.4) vsurf_ID6 (14.5) PEOE_VSA-2 (11) vsurf_DW23 (10.3) **pMRP1 (9.9)** vsurf_DD23 (9.1) Num_Rings_6 (7.8) Num_Rings_5 (6) **pOATP1B1 (3.8)** reactive (3.1) Num_Rings_4 (3.1) Num_Rings_3 (1.2) **ePL (0.2)**	FiA (60.3) LogD(10) (51.4) FiB (43.9) ASA_P (36) LogD(7_4) (35.9) SMR_VSA0 (32.4) vsurf_HL1 (32) Hetero_ratio (31.8) SlogP_VSA5 (29.3) b_1rotR (28.6) MNDO_HOMO (27.6) vsurf_CW6 (26.6) LogP (25.8) PEOE_VSA-0 (24.8) PEOE_PC- (23.6) vsurf_HB2 (23.6) VAdjEq (22.1) vsurf_HB1 (21.9) **pPEPT1 (21.9)** vsa_acc (21.8) density (21) LogD(5_5) (19.6) vsurf_ID4 (19.1) **pBCRP1 (19.1)** Surface_Tension (19) vsurf_ID5 (18.9) **pPL (18.5)** vsurf_EWmin3 (18.2) MNDO_LUMO (17.7) vsurf_HB3 (17.6) SlogP_VSA1 (17.5) vsurf_ID6 (15.9) vsurf_W7 (15.7) vsurf_IW8 (15.4) dens (14.4) AM1_dipole (14) **pMRP2 (13.2)** vsurf_DW23 (11.6) **pMRP1 (10.3)** vsurf_DD23 (9.9) PEOE_VSA-2 (9.7) Num_Rings_6 (7.8) Num_Rings_5 (6.1) reactive (3.5) Num_Rings_4 (3.4) **pOATP1B1 (2.9)** Num_Rings_3 (0.9)

No Feature Selection

Table A3.7. Models based on no feature selection and all features are directly available for training of the QSAR models.

	Random Forest		Boosted Trees	
	ePL (9)	pPL (9a)	ePL (20)	pPL (20a)
MAE	0.318	0.322	0.319	0.316
	ePL (0.1) pMDR1 (3.5) pBCRP1 (3.2) pMRP2 (3.2) pMRP1 (3.2) pOCT1 (2.7) pOATP2B1 (1) pOATP1A2 (3.8) pPEPT1 (5.5) pOATP1B1 (1) pOATP1B3 (4) + All remaining MDs	**pPL (3.5)** pMDR1 (4.2) pBCRP1 (3.4) pMRP2 (3) pMRP1 (2.6) pOCT1 (2.3) pOATP2B1 (0.9) pOATP1A2 (3.6) pPEPT1 (5.4) pOATP1B1 (0.8) pOATP1B3 (4.6) + All remaining MDs	**ePL (0.1)** pMDR1 (3.2) pBCRP1 (3) pMRP2 (3.3) pMRP1 (3.6) pOCT1 (3) pOATP2B1 (0.9) pOATP1A2 (4.1) pPEPT1 (5.5) pOATP1B1 (0.8) pOATP1B3 (4.9) + All remaining MDs	**pPL (3.1)** pMDR1 (4) pBCRP1 (3.5) pMRP2 (4.7) pMRP1 (3) pOCT1 (2) pOATP2B1 (0.6) pOATP1A2 (4.4) pPEPT1 (4.8) pOATP1B1 (2.1) pOATP1B3 (4.2) + All remaining MDs

Table A3.8. Models based on no feature selection and just physiological features available for modelling.

	Random Forest		Boosted Trees	
	ePL (10)	pPL (10a)	ePL (20)	pPL (21a)
MAE	0.380	0.371	0.413	0.412
	pPEPT1 (86.8) pOATP1B3 (77.2) pMDR1 (72.8) pMRP2 (72) pMRP1 (71.4) pBCRP1 (68.1) pOCT1 (65.2) pOATP1A2 (63.2) pOATP2B1 (50.3) pOATP1B1 (33.9) **ePL (3.9)**	pPEPT1 (82.4) **pPL (73.9)** pMRP2 (71.4) pOATP1B3 (71.3) pMDR1 (67) pMRP1 (66.5) pBCRP1 (61.6) pOCT1 (57.9) pOATP1A2 (57.4) pOATP2B1 (43.1) pOATP1B1 (30.5)	pPEPT1 (93.8) pMRP2 (88.8) pMDR1 (78.4) pOATP1B3 (77.3) pBCRP1 (69.5) pOATP1A2 (65.3) pOATP2B1 (62.5) pMRP1 (58.4) pOCT1 (54.1) pOATP1B1 (52.3) **ePL (2.3)**	pPEPT1 (91.9) pMRP2 (85) pPL (77.5) pMDR1 (73.9) pOATP1B3 (72.7) pMRP1 (61.1) pBCRP1 (60.9) pOATP1A2 (59.6) pOATP2B1 (55.3) pOCT1 (49.5) pOATP1B1 (48.6)

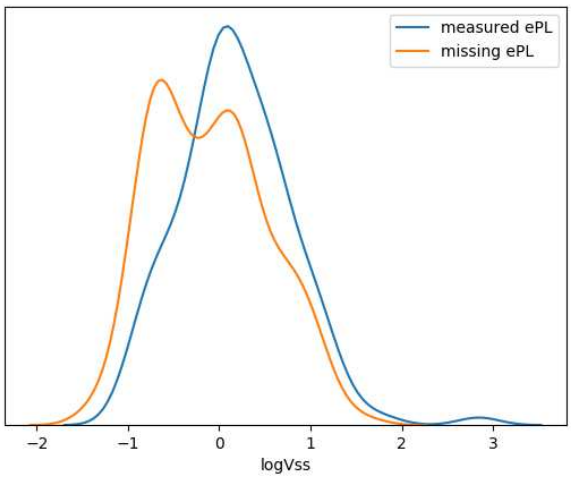


Figure A3.1. Coverage of logVss values from missing phospholipidosis data compared to measured data.

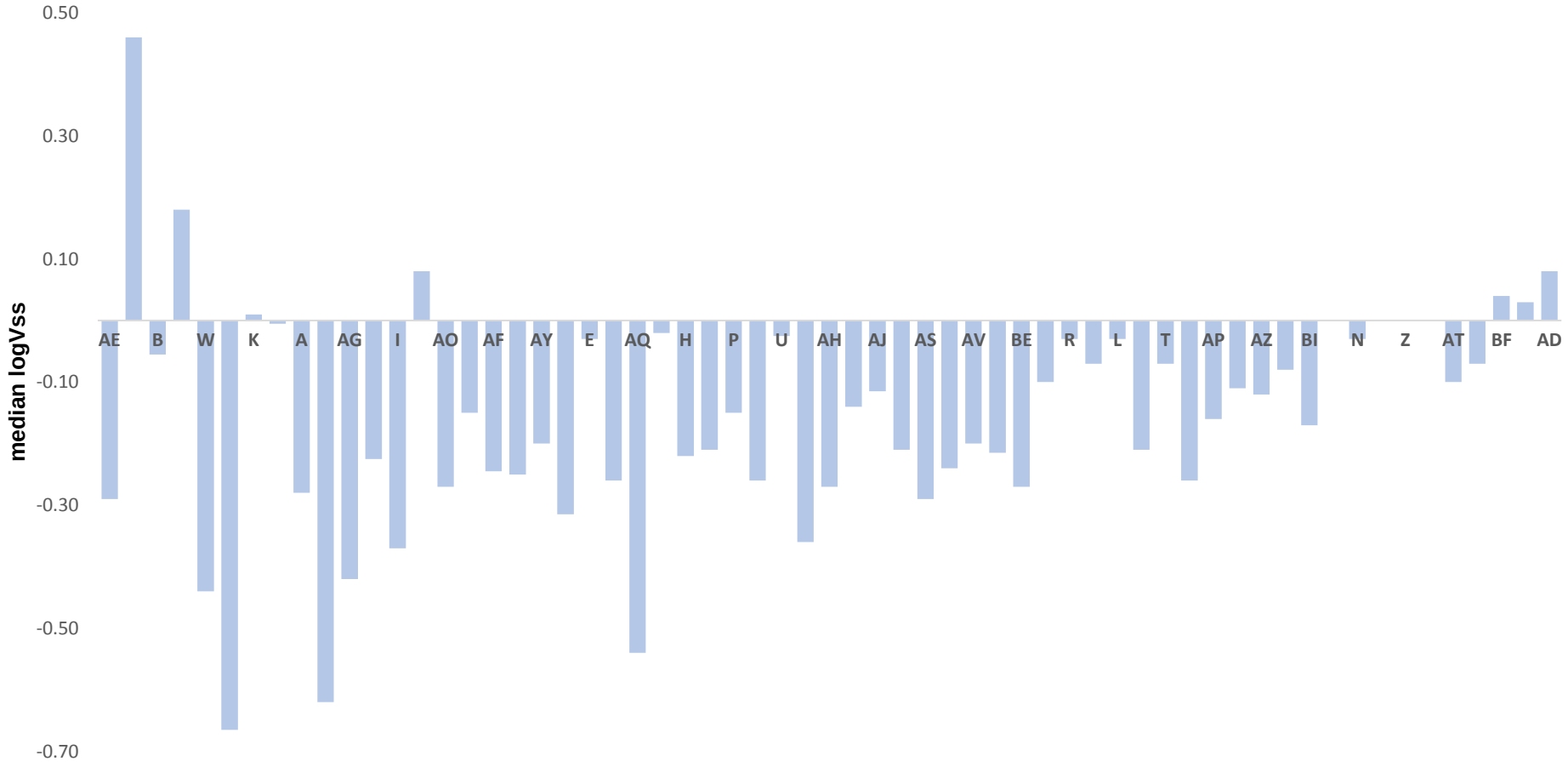


Figure A3.2. Physiological descriptors' combinations sorted in ascending amount of rules associated with each combinations.

Table A3.9. Top node features across the 8a RF model (best overall model).

top node	Number of Trees
FiA	145
LogD(10)	102
Hetero_ratio	84
LogD(7.4)	83
ASA_P	60
vsurf_HL1	60
vsurf_CW6	52
vsurf_HB3	49
SMR_VSA0	48
density	43
vsurf_HB1	42
FiB	34
vsurf_HB2	32
pMRP2	26
Surface_Tension	21
pMRP1	20
dens	16
vsurf_W7	15
pPEPT1	10
LogD(5_5)	9
PEOE_PC-	7
Num_Rings_4	6
PEOE_VSA-2	6
SlogP_VSA1	6
vsurf_EWmin3	4
LogP	4
PEOE_VSA-0	3
AM1_dipole	2
vsurf_ID4	2
vsa_acc	2
MNDO_HOMO	1
vsurf_IW8	1
MNDO_LUMO	1
b_1rotR	1
pPL	1
SlogP_VSA5	1
pOATP1B1	1

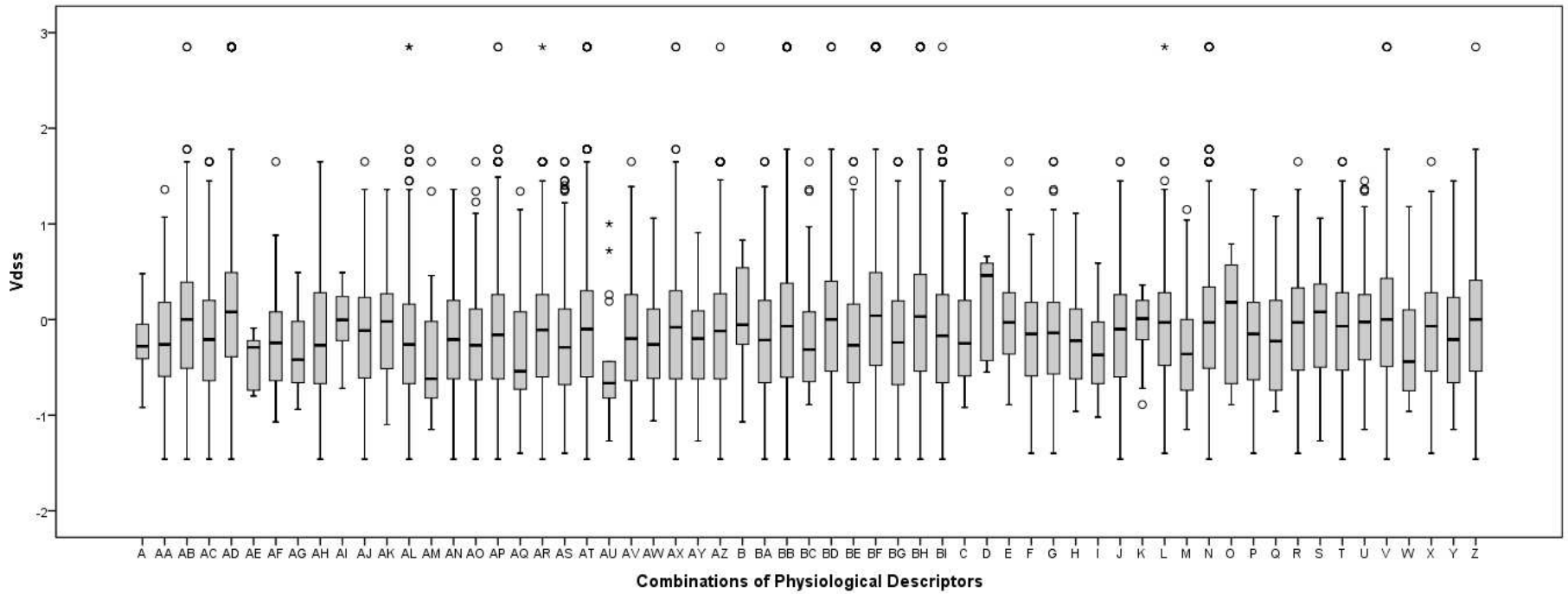


Figure A3.3. Vss distribution across the different combinations of physiological descriptors. Note that a given combination can be found in several different rules of the RF model. The legend of the combinations of physiological descriptors is provided in Table A3.10. This is ordered alphabetically for ease of consultation. Each if-then rule containing these combinations may or may not also contain molecular descriptors.

Table A3.10. Distribution of Vss within each combination, reported in mean, median, standard deviation (SD) and maximum range. This is ordered alphabetically for ease of consultation. Each if-then rule containing these combinations may or may not also contain molecular descriptors. All descriptors listed here are of the type "pDescriptor", but were represented with just the name of the physiological property they encode, for simplicity.

combination code	content	N (rules)	Mean logVss	Median logVss	SD logVss	Range logVss
A	PEPT1, MRP1, MRP2, PL, BCRP1	23	-0.2513	-0.28	0.36199	1.4
AA	PEPT1, BCRP1, OATP1B1	180	-0.1733	-0.26	0.53447	2.82
AB	PEPT1, BCRP1	5184	-0.0048	0	0.59479	4.31
AC	PEPT1, OATP1B1	926	-0.1615	-0.21	0.58675	3.11
AD	PEPT1	25865	0.0755	0.08	0.62438	4.31
AE	MRP1, MRP2, PL, BCRP1, OATP1B1	5	-0.428	-0.29	0.32105	0.71
AF	MRP1, MRP2, PL, BCRP1	96	-0.2328	-0.245	0.46855	2.72
AG	MRP1, MRP2, PL, OATP1B1	37	-0.353	-0.42	0.38781	1.43
AH	MRP1, MRP2, PL	354	-0.162	-0.27	0.63693	3.11
AI	MRP1, MRP2, BCRP1, OATP1B1	22	-0.0527	-0.005	0.36388	1.21
AJ	MRP1, MRP2, BCRP1	496	-0.1317	-0.115	0.56779	3.11
AK	MRP1, MRP2, OATP1B1	139	-0.0432	-0.02	0.54272	2.46
AL	MRP1, MRP2	1885	-0.21	-0.26	0.58532	4.31
AM	MRP1, PL, BCRP1, OATP1B1	33	-0.3603	-0.62	0.66254	2.8
AN	MRP1, PL, BCRP1	519	-0.1597	-0.21	0.54979	2.82
AO	MRP1, PL OATP1B1	69	-0.1523	-0.27	0.64937	3.11
AP	MRP1, PL	2251	-0.1167	-0.16	0.61284	4.31
AQ	MRP1, BCRP1, OATP1B1	115	-0.309	-0.54	0.60116	2.74
AR	MRP1, BCRP1	2405	-0.0979	-0.11	0.59453	4.31
AS	MRP1, OATP1B1	526	-0.2159	-0.29	0.58431	3.05
AT	MRP1	11151	-0.0682	-0.1	0.65122	4.31
AU	MRP2, PL, BCRP1, OATP1B1	18	-0.4867	-0.665	0.62222	2.27
AV	MRP2, PL, BCRP1	653	-0.1288	-0.2	0.59367	3.11
AW	MRP2, PL, OATP1B1	112	-0.2079	-0.26	0.48515	2.12
AX	MRP2, PL	3111	-0.0927	-0.08	0.61257	4.31
AY	MRP2, BCRP1, OATP1B1	98	-0.1972	-0.2	0.47282	2.18
AZ	MRP2, BCRP1	2986	-0.1147	-0.12	0.60462	4.31
B	PEPT1, MRP1, MRP2, PL, OATP1B1	10	-0.017	-0.055	0.63239	1.9
BA	MRP2, OATP1B1	658	-0.1965	-0.215	0.56452	3.11
BB	MRP2	14415	-0.055	-0.07	0.65701	4.31
BC	PL, BCRP1, OATP1B1	104	-0.2414	-0.315	0.50817	2.54
BD	PL, BCRP1	3489	-0.0053	0	0.60572	4.31
BE	PL, OATP1B1	686	-0.2083	-0.27	0.58201	3.11
BF	PL	16897	0.0435	0.04	0.64482	4.31
BG	BCRP1, OATP1B1	580	-0.1749	-0.24	0.59902	3.11
BH	BCRP1	17439	0.0109	0.03	0.63967	4.31
BI	OATP1B1	3202	-0.1269	-0.17	0.64621	4.31
C	PEPT1, MRP1, MRP2, PL	97	-0.1876	-0.25	0.49824	2.03
D	PEPT1, MRP1, MRP2, BCRP1, OATP1B1	5	0.146	0.46	0.58654	1.21
E	PEPT1, MRP1, MRP2, BCRP1	105	-0.037	-0.03	0.51635	2.54
F	PEPT1, MRP1, MRP2, OATP1B1	69	-0.1549	-0.15	0.50079	2.29
G	PEPT1, MRP1, MRP2	474	-0.1349	-0.14	0.52347	3.05
H	PEPT1, MRP1, PL, BCRP1	140	-0.2095	-0.22	0.45953	2.07
I	PEPT1, MRP1, PL, OATP1B1	39	-0.3395	-0.37	0.39512	1.61
J	PEPT1, MRP1, PL	707	-0.1075	-0.1	0.56732	3.11

Table A3.10. (cont.)

combination code	content	N (rules)	Mean logVss	Median logVss	SD logVss	Range logVss
K	PEPT1, MRP1, BCRP1, OATP1B1	18	-0.0783	0.01	0.3731	1.25
L	PEPT1, MRP1, BCRP1	908	-0.0527	-0.03	0.55647	4.25
M	PEPT1, MRP1, OATP1B1	205	-0.3257	-0.36	0.50459	2.3
N	PEPT1, MRP1	3937	-0.0151	-0.03	0.60731	4.31
O	PEPT1, MRP2, PL, BCRP1, OATP1B1	13	-0.0646	0.18	0.66988	1.68
P	PEPT1, MRP2, PL, BCRP1	176	-0.1866	-0.15	0.51236	2.76
Q	PEPT1, MRP2, PL, OATP1B1	38	-0.1939	-0.225	0.53364	2.04
R	PEPT1, MRP2, PL	759	-0.0468	-0.03	0.54531	3.05
S	PEPT1, MRP2, BCRP1, OATP1B1	47	-0.0515	0.08	0.57722	2.33
T	PEPT1, MRP2 BCRP1	1093	-0.0777	-0.07	0.52804	3.11
U	PEPT1, MRP2 OATP1B1	182	-0.0334	-0.025	0.5494	2.6
V	PEPT1, MRP2	4606	0.0058	0	0.60026	4.31
W	PEPT1, PL, BCRP1, OATP1B1	15	-0.2393	-0.44	0.68706	2.14
X	PEPT1, PL, BCRP1	844	-0.0663	-0.07	0.54892	3.05
Y	PEPT1, PL, OATP1B1	161	-0.1498	-0.21	0.60702	2.6
Z	PEPT1, PL	5020	-0.0045	0	0.60508	4.31

Figure A3.4. Predictions of the Gombar external test set.

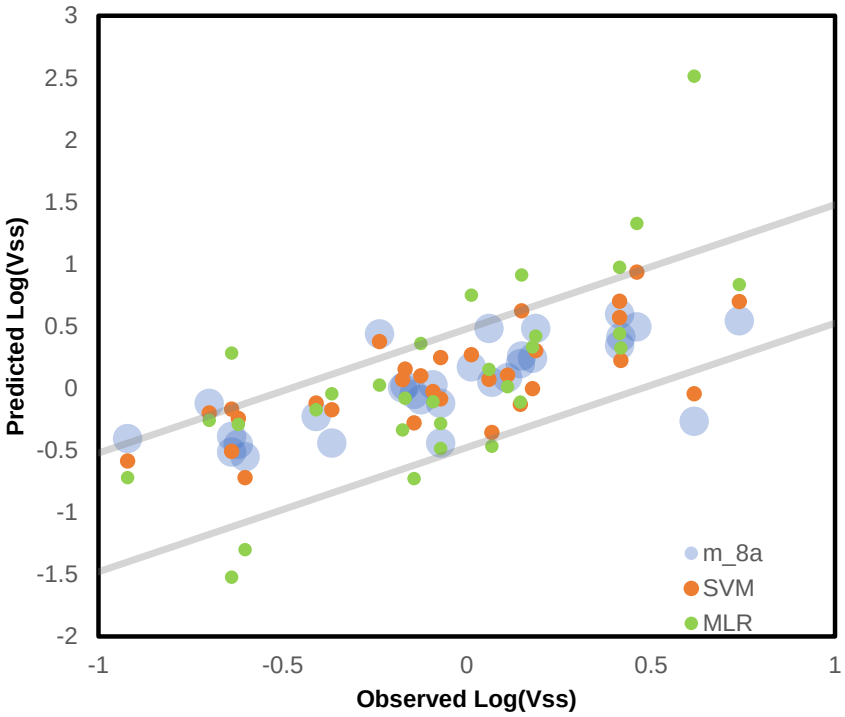
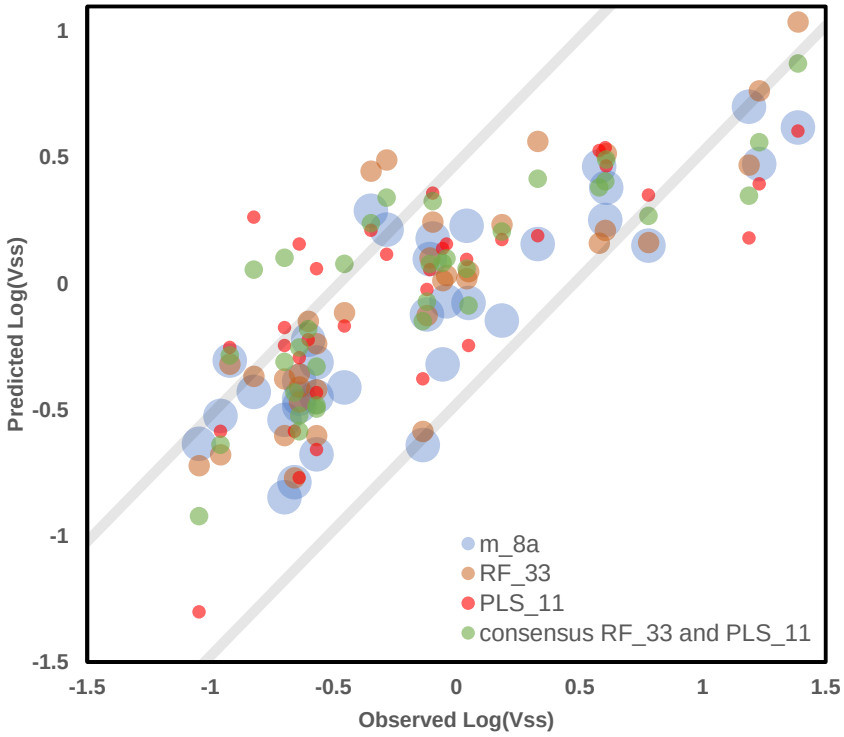


Figure A3.5. Predictions of the Lombardo external test set.



11.4. Appendix IV: Supporting Information for Chapter 7

Table A4.1. Full list of descriptors and their feature importance (% correctly learned compounds) in model 8a (the best model in our previous work).

MAE (IV set) = 0,306	
FiA (47.1)	Surface_Tension (18.4)
LogD(10) (38.5)	SlogP_VSA1 (18.4)
LogD(7_4) (31.9)	**pBCRP1 (17.7)**
FiB (30.8)	MNDO_LUMO (17.2)
ASA_P (30.6)	**pPL (17.1)**
Hetero_ratio (29.6)	vsurf_EWmin3 (17)
vsurf_HL1 (29)	vsurf_ID5 (16.9)
vsurf_CW6 (25.3)	dens (16.5)
vsurf_HB1 (24.9)	vsurf_W7 (16.3)
pPEPT1 (24.5)	vsurf_IW8 (15.5)
vsurf_HB2 (24.2)	**pMRP2 (15.3)**
density (24)	AM1_dipole (15.3)
SMR_VSA0 (23.4)	vsurf_ID6 (14.9)
MNDO_HOMO (23.4)	**pMRP1 (12.5)**
vsurf_HB3 (22.6)	vsurf_DD23 (11.4)
PEOE_VSA-0 (22.6)	vsurf_DW23 (10.9)
LogP (22.5)	PEOE_VSA-2 (9.6)
b_1rotR (21.1)	Num_Rings_6 (9.5)
SlogP_VSA5 (20.6)	Num_Rings_5 (6.1)
VAdjEq (20.2)	**pOATP1B1 (3.8)**
PEOE_PC- (19.9)	reactive (3.6)
vsa_acc (19.9)	Num_Rings_4 (3.3)
LogD(5_5) (19.3)	Num_Rings_3 (1.3)
vsurf_ID4 (18.9)	

Table A4.2. Full list of combinations in the best model (M5) retrained in the presence of pPPB, and the full list of combinations in M5.

M5 retrained with pPPB	M5
pPPB	OATP1B1_c
pBCRP1_c	BCRP1_c
pMRP2_c	MRP2_c
pBCRP1_c + pPPB	BCRP1_c + pOATP1B1_c
pMRP2_c + pPPB	MRP2_c + pOATP1B1_c
pOATP1B1_c	BCRP1_c + pMRP2_c
pMRP2_c + pBCRP1_c	BCRP1_c + pMRP2_c + pOATP1B1_c
pOATP1B1_c + pPPB	
pBCRP1_c + pOATP1B1_c	
pMRP2_c + pBCRP1_c + pPPB	
pMRP2_c + pOATP1B1_c	
pMRP2_c + pOATP1B1_c + pPPB	
pMRP2_c + pBCRP1_c + pOATP1B1_c	
pBCRP1_c + pOATP1B1_c + pPPB	

11.5. Appendix V: Supporting Information for Chapter 8

Supplement A5.1: Impact of the minimum required number of training neighbours

The results presented regarding RDN consider an external compound within the AD if it falls within the threshold distance of at least 1 training compound at any given iteration of the algorithm (refer to scheme 1, where at the last step new instances will be considered as being covered if falling within "any $MaxDist_i$", meaning within at least 1 training neighbourhood). To explore the impact of this parameter, the effect of increasing the minimum number of required nearest neighbours was tested. Except for a required minimum of 2 nearest neighbours (2 NN_{min}), increasing the number of training neighbours revealed to be useless, yielding a low quality AD core often worse than the baseline accuracy achieved when all data are considered. Imposing a restriction of 2 NN_{min} showed higher quality at the inner most region of the model (Figure SI1) compared to when one single neighbour is required (1 NN_{min}) (Figure 8, Results and Discussion), but on the other hand the obtained profiles from the latter were smoother.

As a result, it is not straightforward to choose one alternative over the other. However, this experiment showed that, counterintuitively, having 1 neighbour as minimum requirement does not only provide a useful AD but it is better than, say, 4 nearest neighbours. This is in line with our remaining observations that point towards the importance of addressing small regions in the chemical landscape.

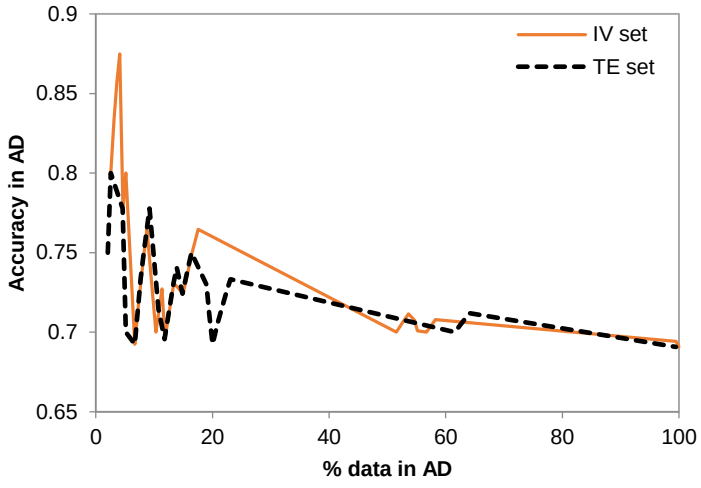


Figure A5.1. RDN AD with minimum 2 nearest neighbours required in order for a query to be considered included in the AD.

Supplement A5.2: Complementary assessment of simpler curve similarity measures

To complement the analysis of the AD scoring function other simpler measures were analysed. A pairwise similarity was calculated based on comparing every sub-section between the curves for the two external subsets, and counting the percentage of matching segments (in terms of slope) between a pair of curves (Table A5.1). In all datasets this measure produces at least one occasion where higher pairwise similarity does not correspond to a visually better profile. This could be explained by the fact that purely looking at the similarity between curves does not distinguish between descending and increasing trends, which led us to conclude that absolute similarity in itself is insufficient in assessing the quality of an AD profile. Another evaluated measure, which is a part of the AD scoring function, is the SMP. This could be considered as a more sophisticated pairwise similarity, as it takes into account both slope mismatch and slope direction (results summarized in Table A5.2). In this case, the lower the value, the better is the overall trend between both external curves. The average SMP is consistently lower with STD across all datasets, however as already explained this could be misleading as matching slopes between a pair of curves can have a different value according to the amount of data associated with each section.

Lastly, the Area Between Curves (ABC) (Table SI2) shows that all three datasets have the smallest ABC with dk-NN, which again shows this would be misleading to use as an assessment measure. Solely having a small absolute difference between both curves does not necessarily mean the AD profile has more quality, as shown by Ames dk-NN AD where the two curves are close to each other, yet they have a very poor characterization of the model's AD.

Table A5.1. Pairwise similarity across all three models studied. Pairwise Similarity indicates the percentage of segments in both external set curves which show a matching slope. The best value in each dataset is highlighted in bold.

	PAIRWISE SIMILARITY (%)			
	RDN	STD	dk-NN	KDE
P-GP	50	50	**68**	55
AMES	45	**71**	46	64
CYP450	55	**88**	46	63

Table A5.2. Summary of the average Slope Mismatch Penalty (SMP) and the Area Between Curves (ABC) across all three models studied. The SMP was calculated as explained in the methods section, and the ABC was calculated from the sum of the area of trapezoids formed between the two curves under analysis. The best value in each dataset is highlighted in bold.

AVERAGE SMP	**AREA BETWEEN CURVES (ABC)**

	RDN	STD	dk-NN	KDE	RDN	STD	dk-NN	KDE
P-GP	6.8	**3.6**	4.6	7.2	1.9	1.9	**0.05**	1.7
AMES	6.7	**3.7**	5.5	8.1	1.9	0.4	**0.1**	0.9
CYP450	7.1	**3.0**	7.3	9.2	0.5	**0.4**	0.1	1.1