DATA MINING METHODS FOR THE PREDICTION OF INTESTINAL ABSORPTION USING QSAR

A THESIS SUBMITTED TO

THE UNIVERSITIES OF KENT AND GREENWICH AT MEDWAY

IN THE SUBJECT OF PHARMACY

FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY

By

Danielle Anne Newby

November 2014

Abstract

Oral administration is the most common route for administration of drugs. With the growing cost of drug discovery, the development of Quantitative Structure-Activity Relationships (QSAR) as computational methods to predict oral absorption is highly desirable for cost effective reasons. The aim of this research was to develop QSAR models that are highly accurate and interpretable for the prediction of oral absorption. In this investigation the problems addressed were datasets with unbalanced class distributions, feature selection and the effects of solubility and permeability towards oral absorption prediction. Firstly, oral absorption models were obtained by overcoming the problem of unbalanced class distributions in datasets using two techniques, under-sampling of compounds belonging to the majority class and the use of different misclassification costs for different types of misclassifications. Using these methods, models with higher accuracy were produced using regression and linear/non-linear classification techniques. Secondly, the use of several pre-processing feature selection methods in tandem with decision tree classification analysis – including misclassification costs – were found to produce models with better interpretability and higher predictive accuracy. These methods were successful to select the most important molecular descriptors and to overcome the problem of unbalanced classes. Thirdly, the roles of solubility and permeability in oral absorption were also investigated. This involved expansion of oral absorption datasets and collection of *in vitro* and aqueous solubility data. This work found that the inclusion of predicted and experimental solubility in permeability models can improve model accuracy. However, the impact of solubility on oral absorption prediction was not as influential as expected. Finally, predictive models of permeability and solubility were built to predict a provisional Biopharmaceutic Classification System (BCS) class using two multi-label classification techniques, binary relevance and classifier chain. The classifier chain method was shown to have higher predictive accuracy by using predicted solubility as a molecular descriptor for permeability models, and hence better final provisional BCS prediction. Overall, this research has resulted in predictive and interpretable models that could be useful in a drug discovery context.

Acknowledgments

I would firstly like to thank both my supervisors Dr Taravat Ghafourian and Professor Alex Freitas for all of their advice and guidance on this project; without their continued enthusiasm, help and support, this research would not have been possible.

To my mum, dad and Pippa – you have supported me through all the years, and thanks for putting up with me while I ate, slept and breathed my PhD!

Many thanks to Liz and Jeremy Francis and all their family (including the cats Molly, Gizmo and Tilly!), for making me feel so welcome when I first moved down to Medway when I started my PhD.

A special thank you to Shelagh (and Maggie!), thank you for all your help, support and occasional walking breaks!

My PhD would not have been the same without the amazing group of people who studied alongside me throughout my time at Medway, so a big thank you for making the past three years an enjoyable experience. A special mention has to go to Lynsey Atkinson, you are a true friend, and thank you for keeping me sane throughout my PhD!

Finally, to Sam, thank you for being there, for your love and constant support – I would have not been able to do this without you by my side! -

Contents

List of Figures

List of Tables

List of Key Abbreviations

%HIA	Percentage human intestinal absorption
ADMET	Absorption, Distribution, Metabolism, Excretion and Toxicity
BCS	Biopharmaceutics classification system
C&RT	Classification and Regression Trees
Caco-2	Human colon adenocarcinoma
CS	Chi Square
DT	Decision Trees
F	Bioavailability
FA	Fraction absorbed/Oral absorption
GEN	Genetic Algorithm
GIT	Gastrointestinal tract
GSE	General Solubility Equation
IGR	Information Gains Ratio
LDA	Linear Discriminant Analysis
MDCK	Madin-Darby Canine Kidney
MLR	Multiple Linear Regression
MPbAP	Melting point based absorption potential
NCE	New Chemical Entities
PK	Pharmacokinetics
PSA	Polar Surface Area
QSAR	Quantitative Structure-Activity Relationship
RF	Random Forest
RMSE	Root Mean Squared Error
SVM	Support Vector Machine
T set	Training set
V set	Validation set

1. Introduction

The cost of bringing a drug onto the market keeps on rising (DiMasi *et al*., 2003). Therefore drug companies look for methods to increase cost effectiveness. There is a demand for drugs with good absorption, distribution, metabolism and excretion (ADME) properties. Originally these properties were only tested later on in drug development; however this resulted in a high attrition rate due to poor pharmacokinetics. By testing ADME properties in early drug discovery with experimental and/or *in silico* models there has been a reduction in failure rates due to poor pharmacokinetics (Kola and Landis, 2004). Even with the increase in automation, high throughput screening experimental assays are suffering bottlenecks in drug discovery due to the vast number of compounds. *In silico* models, particularly quantitative structure activity relationships (QSAR), can predict ADME properties of new compounds from molecular descriptors calculated just from chemical structure. QSAR models offer an appealing cost effective addition or alternative to remove compounds with undesirable properties as early as possible without chemical synthesis or testing (Yu and Adedoyin, 2003).

Oral administration is the most common and popular route of administration. For that reason, the accurate prediction of oral absorption is highly desirable. However, oral absorption depends on many physiological, physicochemical and formulation factors, making prediction by QSAR models a challenge. In addition, high quality data are required with relevant molecular descriptors that can produce models that try to take in account all the mitigating drug related factors affecting absorption (Hou *et al*., 2007c). The focus of any QSAR model, including an oral absorption model, is to be predictive and suitable for intended use, yet interpretability of the model is also desirable.

Thus, this work starts by branching out from previous research carried out on published oral absorption datasets. The publication of a large dataset of over 600 compounds offers a good starting point for building models to predict oral absorption (Hou *et al*., 2007c). The main problem with previous oral absorption models is that they are built using datasets that contain a much higher proportion of

highly-absorbed compounds compared to poorly-absorbed compounds. This results in models biased towards the prediction of highly-absorbed compounds and not reflective of a drug discovery scenario where there are more poorly-absorbed than highly-absorbed compounds (Lipinski, 2000).

In many applications of QSAR, there are a large variety of molecular descriptors utilised to build models. The use of too many descriptors can result in over-fitting, lower predictability and a decrease in interpretability of resulting models (Goodarzi *et al*., 2012). Therefore exploration of feature selection methods in relation to oral absorption can be investigated.

With the lack of oral absorption data in recent publications, this work expands by focussing on the two main factors influencing absorption: permeability and solubility, which are measured frequently in drug discovery. The lack of oral absorption models which take into account these properties, in particular solubility, is not reflective of an industry scenario where there are an increasing number of poorly soluble compounds (Williams *et al*., 2013). Therefore, the collection of permeability and solubility data from the literature to study the influence of these two properties on oral absorption is explored in this research.

Finally, there is a growing number of models that predict permeability and solubility separately and fail to take into account the interaction between these two properties (Gozalbes *et al*., 2011, Gozalbes and Pineda-Lucena, 2010). Therefore, models which predict these two properties simultaneously and take into account the interaction are also examined in this research.

1.1 Aims and Objectives

The ultimate aim of this work is to develop models that are predictive and interpretable for a drug discovery context. Different approaches will be explored in this project in order to produce oral absorption models that achieve the project's aims, which are summarised in Figure 1.1

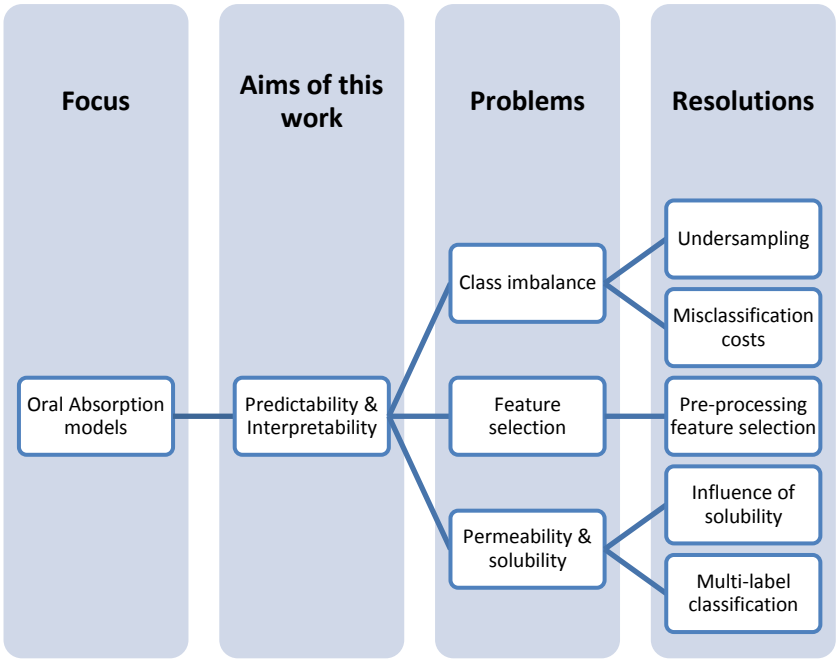

Figure 1.1. Overview of the development of interpretable and predictable oral absorption models

As shown in Figure 1.1, this work will follow the three problems and the resolutions to these problems in order to accomplish the aims of this work, to produce models that are predictive and interpretable.

Specific aims are:

- Investigate the effects of under-sampling and misclassification costs to overcome the problem of datasets with unbalanced class distribution and improve model accuracy, this was presented in chapter 6 and 7;
- Determine the influence of pre-processing feature selection methods in the interpretation and predictive ability of oral absorption models as shown in chapter 8;
- Examine the effect of solubility and permeability for oral absorption prediction as defined in chapter 9;
- Compare multi-label classification methods to predict two rate limiting steps of absorption: solubility and permeability, for provisional Biopharmaceutics Classification System (BCS) class prediction in the final experimental chapter 10.

The experimental chapters that deal with the specific aims of this thesis can be found in chapters 6-10. The final chapters present a summary of the key findings (chapter 11) and discuss potential future research directions related to this project (chapter 12). In addition, Appendices 1-4 provide additional information relating to the results of the experimental chapters 6-10. The datasets used in this work and extra supporting information which was too large to include in the appendices can be found with the accompanying disk with this thesis.

1.2 Original Contributions

The summary of the main contributions of this thesis is presented below:

Firstly, this is the first work to my best knowledge which presents two methods for overcoming the problem of oral absorption datasets with unbalanced class distribution and compares them with models based on these data unbalanced datasets using a variety of linear and non-linear data mining methods, predicting categorical and numerical variables. These methods included training set selection by under-sampling of the majority class and the use of higher misclassification for the classification of the minority class. The methods for overcoming unbalanced datasets were further extended to showing the influence of higher misclassification costs on a balanced dataset where the majority class was under-sampled for the prediction of oral absorption.

Secondly, the combination of various pre-processing feature selection methods with misclassification costs has not been carried out for oral absorption models in the literature, to the best of my knowledge. Therefore, in this work, the effects of various pre-processing (as opposed to embedded) feature selection methods in conjunction with the use of various misclassification costs were examined.

The expansion of the fraction absorbed dataset with more drugs, including the collection of permeability, solubility, melting point and maximum dose data could be useful for those interested in modelling this type of data for drug and drug like properties. From this dataset, this work presents the inclusion of experimental permeability and solubility in oral absorption models to see the influence (if any) they have on oral absorption prediction.

Finally, based on the collected data and published dataset, this research offers a first use of the classifier chain method for the multi-label classification of permeability and solubility as a provisional biopharmaceutics classification system (BCS) class prediction suitable for drug discovery.

From the above contributions the aims of this thesis to obtain interpretable and predictable models have been achieved, as will be shown by the computational results reported later in this thesis.

!
!
!
!
!
!
!
!
!
!
!
!
!
!
!

2. Drug Discovery and Oral Absorption

2.1 Relevance to the Pharmaceutical Industry

The cost to bring a drug to market is rapidly increasing. The time taken to successfully progress a drug from discovery to market is on average 10 years with costs of over 1 billion dollars (DiMasi *et al*., 2003). The high failure rate of drug candidates adds to the pressure to produce a money making blockbuster drug in the shortest time possible (Bunnage, 2011).

Drug discovery has now shifted from primarily focusing on efficacy and selectivity of new drug candidates, to the incorporation of testing of absorption, distribution, metabolism, elimination and toxicity (ADMET) properties. Historically, these properties were usually characterised later on in drug development. The incorporation of testing these ADMET properties earlier has resulted in a decrease in drug candidate failure rate due to poor ADMET properties, from 40% to 10%, in phase I clinical trials (Kola and Landis, 2004). This is one of the many cost effective strategies employed by pharmaceutical companies that can ensure recognition and elimination of unsuitable compounds as early as possible using a "fail fast, fail cheap" approach (Yu and Adedoyin, 2003).

High throughput experimental assays that evaluate ADMET properties are carried out across the main stages of drug discovery (Figure 2.1). Even with the help of high throughput automation, there is still a large bottleneck due to the overwhelming number of potential hits in the lead identification and optimisation (Gleeson *et al*., 2011, Caldwell *et al*., 2001).

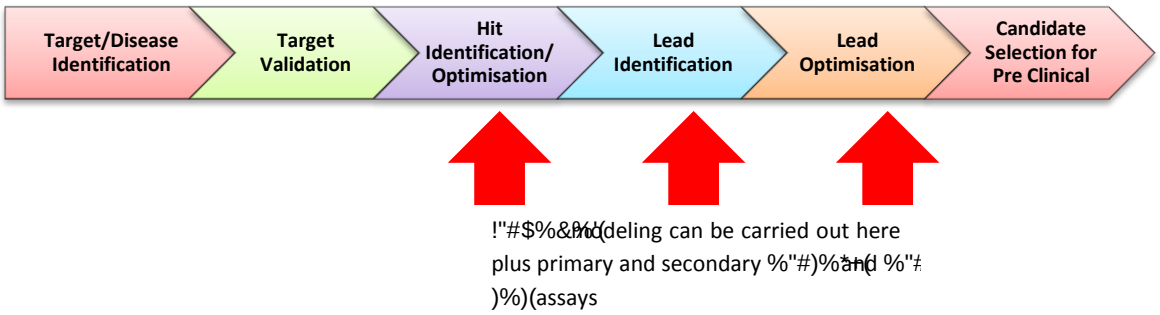

Figure 2.1. Simplified summary of drug discovery process

In silico modelling of ADMET properties (particularly absorption) has been incorporated into the hit to lead screening of compounds in tandem with experimental ADMET assays (Caldwell *et al*., 2009, Gleeson *et al*., 2011). The benefit of using *in silico* models is that predictions can be made based on chemical structure alone with no chemical synthesis. Therefore *in silico* models avoid the experimental screening of potentially millions of compounds providing a fast cost-effective method to solve the bottleneck in lead identification and optimisation (Kortagere and Ekins, 2010, Oprea and Matter, 2004).

Therefore *in silico* models can be used as a cost-effective strategy to remove unsuitable compounds as soon as possible whilst fast tracking promising ones. Additionally, the predictions can act as a guidance tool to help select the next appropriate assays to perform in the drug discovery process (van de Waterbeemd and Gifford, 2003, Geerts and Heyden, 2011).

Progress has been made by incorporating *in silico* modelling into early drug discovery programmes; however, there are always constant improvements that can be made to produce models that are predictive and interpretable that will help to bring a money-making drug to market.

2.2 Pharmacokinetics and ADMET

Pharmacokinetics (PK) is the mathematical description of what the body does to the drug from administration to the site of action (Figure 2.2). This can be further expanded as the study of rates of ADMET of the drug after administration (Kwon, 2002, Rosenbaum, 2011)

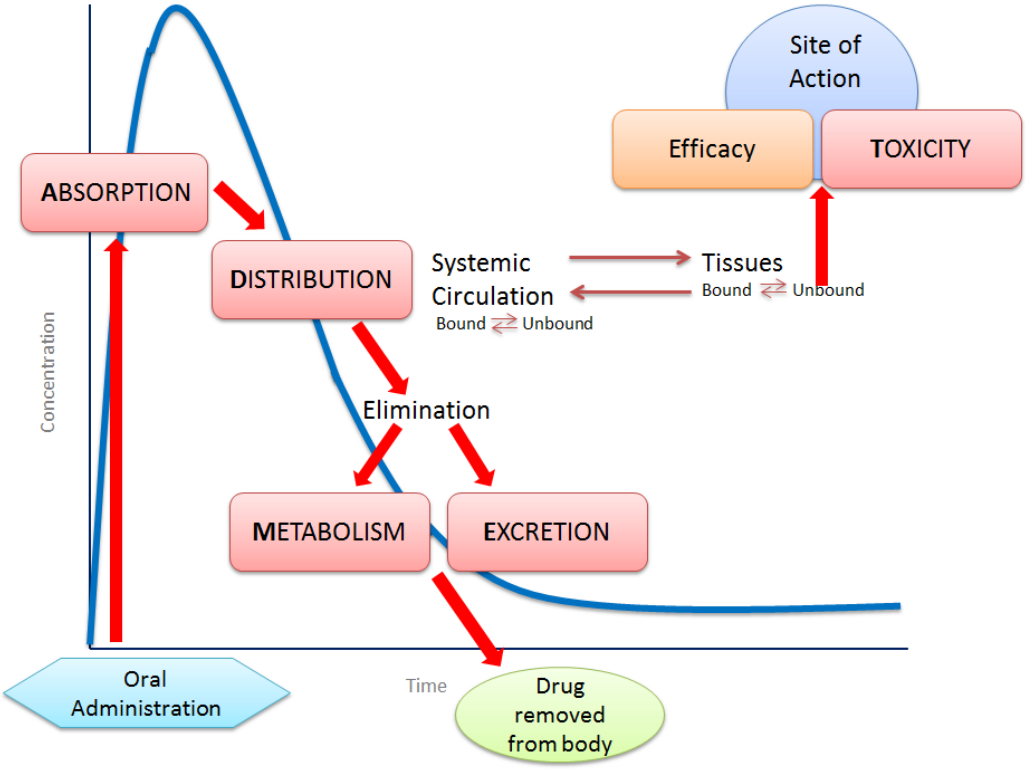

Figure 2.2. The pharmacokinetic profile of a drug after oral administration, indicating the path of the drug from administration to the site of action and removal from the body

Absorption is the first process a drug must overcome from the administration site to reach the systemic circulation. It is influenced by a variety of physiological effects which, in turn will be dependent on the physicochemical properties of the drug and formulation factors. A more in depth discussion of oral absorption is presented in the later sections.

Distribution of a drug shows how well and quickly a drug reversibly transfers between different compartments within the body. The distribution stage is important as compounds must distribute throughout the body, through different tissues, in order to reach the target site and exert a pharmacological effect (Caldwell *et al*., 2009). The rate and extent of distribution for a drug will depend on physicochemical properties such as affinity for particular tissues, tissue composition, tissue volume and protein binding (Dowty *et al*., 2011, Kirkovsky and Zutshi, 2011).

Metabolism refers to the structural modification of the drug into product metabolites by a variety of enzymes in the body. This process occurs mainly in the liver, however, metabolism also occurs in the small intestine, lungs, kidney and other organs. The aim of metabolism is to convert hydrophobic drug compounds into more hydrophilic entities that can be readily excreted and removed from the body. This is carried out by phase I and II metabolic phases governed by different metabolising enzymes (King, 2009).

Excretion is the removal of the unchanged drug or hydrophilic metabolites through the kidneys (urine) or faeces. Hydrophilic compounds tend to be excreted via the kidneys whereas larger lipophilic compounds are excreted into faeces via the bile duct (Caldwell *et al*., 1995, Ghibellini *et al*., 2006). Defining the main route of elimination is important firstly to avoid compound accumulation and potential toxic effects. This is particularly important in cases of hepatic or renal failure. Secondly, the data from these excretion studies can be used to calculate intestinal absorption in humans which can be used for *in silico* modelling (Zhao *et al*., 2001).

Finally, **toxicity** is now included with the other ADME processes as it is important with regards to drug safety. Studies that predict or give information about side effects such as cardiac, genetic and hepatic toxicity can also help guide the selection of drug candidates (Gleeson *et al*., 2012).

There are many factors a drug must overcome and avoid in order to get to the site of action to exert a physiological effect. The common factor in ADMET is that the physicochemical properties of compounds can govern all these processes. Therefore, defining the relationship between the chemical structure and these properties can be a powerful tool to determine which drugs have the best ADMET profiles and are potential marketable drugs.

2.3 Process of Oral Drug Absorption through the Body

In order to appreciate the factors that affect oral absorption it is important to understand the route a drug takes in order to be absorbed.

The gastrointestinal tract (GIT) comprise of four main anatomical areas:

- Oesophagus
- Stomach
- Small intestine
- Large intestine

The overall function of the GIT is to break down materials and move materials such as water, nutrients and electrolytes from the external environment of the body into the internal environment of the body. In addition there are many protective mechanisms which prevent harmful foreign substances (including drugs) from reaching the internal tissues. These physiological factors can determine if a drug will be a successful orally administered compound (Shen, 2009).

When a drug is swallowed, e.g. as a tablet or capsule, it travels down the oesophagus to join into the stomach via the cardiac orifice. The stomach is an acidic reservoir containing gastric acid secreting cells and various enzymes which can degrade, break up and mix the drug in the GIT fluid. The process of gastric emptying delivers drug in the GIT fluid to the small intestine at a controlled rate (Silverthorn, 2001, Pal *et al*., 2007).

The small intestine is divided into 3 parts:

- Duodenum
- Jejunum
- Ileum

The small intestine is the major site for oral absorption of many drugs. However the specific region of absorption will depend on the physicochemical properties of the drug itself as well as formulation, pH and transporter abundance in the small intestine regions (Kay, 2011, Petri and Lennernäs, 2003, Martinez and Amidon, 2002). Only drug that is dissolved in the GIT fluid can be absorbed. The dissolved drug must diffuse through the unstirred water layer (UWL) in order to reach the intestinal cell wall. Once through the UWL, the drug can pass through the gut absorptive cells (enterocytes) by a variety of mechanisms. Once the drug molecules have passed through enterocytes into the hepatic portal vein via the mesenteric

capillary network, the drug is considered absorbed. Once absorbed, other pharmacokinetic processes take affect hopefully leading to the exertion of the pharmacological effect of the drug.

There are many physiological barriers a drug must overcome to be absorbed, additionally there are many more after this to prevent a drug reaching its intended target. The process of oral absorption of a drug needs to be understood from a physiological perspective to develop more robust *in silico* models that offer a mechanistic understanding of the many barriers faced by a drug after oral administration.

2.4 Structure of the Small Intestine

To understand why the small intestine is well adapted for extensive absorption of nutrients and also drug compounds, this section highlights the structural anatomy of the gut wall and the intestinal cells in relation to oral absorption.

The wall of the gastrointestinal tract has 4 fundamental histological layers:

- Serosa
- Muscularis externa
- Sub mucosa
- Mucosa

These four layers are the main features, however there are adaptations in different regions and this enables the GIT to carry out its function of digestion and absorption (Pocock and Richards, 2009). The two inner layers of the sub mucosa and the mucosa are important in relation to drug absorption due to the increase in surface area and substantial blood supply of the small intestine (Figure 2.3).

Firstly, the mucosa has finger-like extensions protruding into the intestinal lumen increasing surface area for potential absorption, and on top of this, each individual villus is covered with microvilli (brush border) to enhance the surface area even further to about $200m^2$. This is a massive surface area for potential drug absorption compared with the stomach which has a surface area of about $1m^2$ (Susanto Park and Chang, 2011, Silverthorn, 2001).

Secondly, the sub mucosa has a substantial blood supply which drives and maintains a concentration gradient. This is important in relation to drug absorption, as the sink conditions promote the drug molecules to leave the enterocyte and not to diffuse back into the cell due to constant removal by the portal vein blood flow and dilution in the blood (Buckley *et al*., 2012).

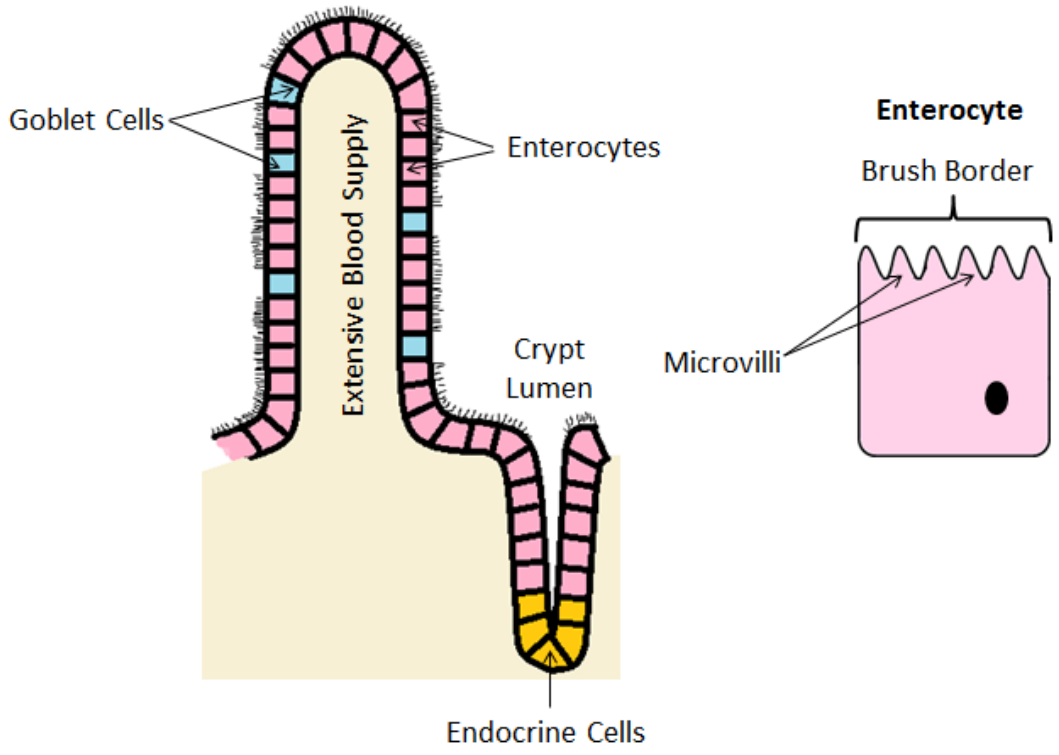

Figure 2.3. Structure of intestinal villi and absorptive cell (enterocyte) showing the microvilli increasing surface area and absorption adapted from (Silverthorn, 2001)

In addition to enterocyte cells, there are also endocrine cells and goblet cells in this layer. Endocrine cells secrete hormones and goblet cells secrete mucus to cover and protect the enterocyte cells. All these cells are created from stem cells in the crypt lumen (Specian and Oliver, 1991). Enterocytes take 3-5 days to differentiate and migrate from the crypt lumen to the villus tip. At the villus tip enterocytes undergo apoptosis and are sloughed off into the intestinal lumen to ensure the small intestine remains efficient and undamaged for effective absorption (Ruemmele *et al*., 2002).

2.5 Routes for Drug Absorption Through the Small Intestine

The main route for drug absorption is controlled by one or more of following mechanisms; transcellular diffusion, paracellular diffusion, efflux and influx through carrier proteins also known as transporters.

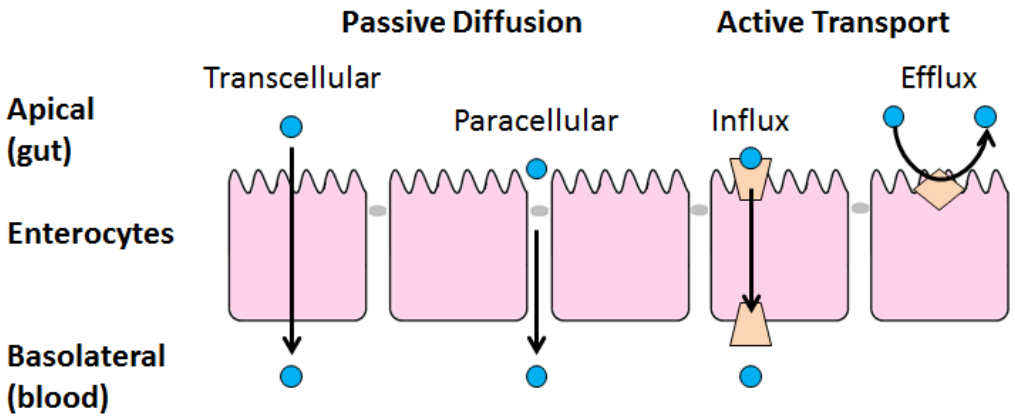

Figure 2.4. Primary mechanisms of intestinal absorption adapted from (Thomas *et al*. 2006)

2.5.1 Passive diffusion

Transcellular and paracellular diffusion can be categorized as passive absorption mechanisms. Passive diffusion is the movement of molecules from an area of high concentration to a low concentration across the cell membrane. This process is governed by Fick's first law (Equation 2.1).

$$\text{Rate of Diffusion} = \frac{' \ \$P \quad \& K \& O}{[} \qquad \textbf{Eq. 2.1}$$

The rate of passive diffusion is dependent on D, the diffusion coefficient which is related to the ability of the drug molecule to diffuse into the cell membrane; A, the surface area of the membrane; P, the partition coefficient relating to the affinity of the drug to the membrane; C*h* – C*l*, is concentration difference between high and low concentrations and finally x, the thickness of the membrane.

Passive diffusion is driven by concentration gradients due to the difference in concentrations across the cell membrane and maintained by sink conditions of the GIT. The concentration gradient for the drug is dependent on GIT physiology, cell

membrane composition and also physicochemical properties of the drug including its solubility and partition coefficients (Singer and Nicholson, 1972, Martinez and Amidon, 2002).

Passive transcellular absorption is where the drug molecule permeates the apical membrane of the enterocyte, diffuses through the cell cytoplasm inside the cell, and finally diffuses out through the basolateral membrane to be absorbed into the blood (Stenberg *et al*., 2000). On the other hand, the passive paracellular route involves compounds being absorbed from between the enterocyte cells through water filled pores and tight junctions. In contrast to the transcellular route, the paracellular route is more selective for small, cationic, hydrophilic drugs (<200 Da and logP < 0) due to the characteristics of the water pores and tight junctions (Petri and Lennernäs, 2003, Martinez and Amidon, 2002). Additionally there is only a small surface area of water pores for paracellular absorption. Due to the increasing tightness of the tight junctions in the later regions of the small intestine this route does not contribute significantly to the intestinal absorption of the vast majority of drugs (Stenberg *et al*., 2000, Ungell *et al*., 1998).

2.5.2 Influx and Efflux Carrier Mediated Transport

Carrier mediated transport involves the movement of molecules across the cell membrane using a transporter protein embedded in the cellular membrane. This active process often requires cellular energy unlike passive absorption which does not. The key to carrier mediated transport is substrate specificity. If a compound has the right specificity for the transporter it will be transported via this route. In the context of absorption, transporters on the apical membrane that can increase drug concentrations into the cell are influx transporters and those that decrease the number of drug molecules entering the cell are termed efflux transporters. Drug design has involved exploiting influx transporter specificity in order to increase the absorption of many drugs (Dobson and Kell, 2008, Kikuchi *et al*., 2009). On the other hand, knowledge regarding substrate specificity of efflux transporters can aid with drug design to avoid making potential substrates for these efflux transporters (Raub, 2005).

There are many differences between passive and carrier mediated transport. One major difference is that carrier mediated transport can be saturated and limited by the abundance of a particular transporter. This can have implications for the oral absorption of compounds which are transport substrates. The impact of transporter saturation can result in non- linear absorption potentially affecting other ADMET processes. The impact of an absorption route can vary depending on the drug; however there is strong evidence to suggest that multiple absorption mechanisms co-exist for the oral absorption for many drugs (Sugano *et al*., 2010, Di *et al*., 2012, Smith *et al*., 2014). However, a counter argument suggests that all absorption is mediated by transporters (Dobson and Kell, 2008, Kell *et al*., 2013). What is clear is that regardless of which absorption mechanism is dominant, carrier mediated transporters can impact on intestinal absorption; therefore increasing research into this area will only benefit understanding. An overview of carrier mediated intestinal transporters is detailed in later sections.

2.6 Intestinal Absorption and Bioavailability

Oral bioavailability and intestinal absorption are important to distinguish as sometimes these terms are used interchangeably. It must be emphasised that intestinal absorption is the prerequisite to oral bioavailability (Zhu *et al*., 2011).

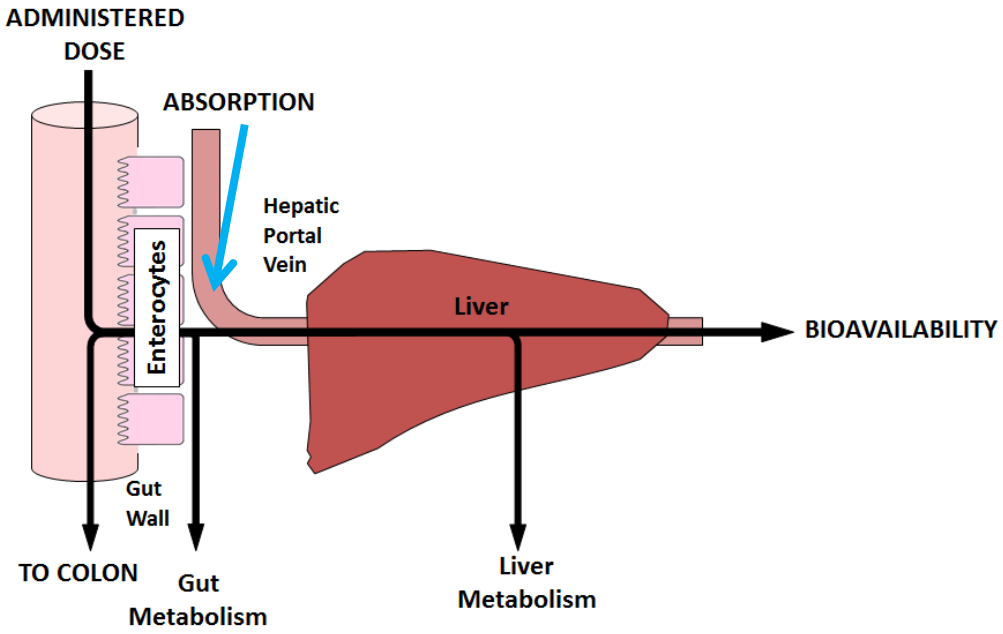

Figure 2.5. The difference between intestinal absorption and bioavailability adapted from (van de Waterbeemd and Gifford 2003)

Intestinal absorption is defined as the amount of drug that passes through the intestinal tissue and enters the portal vein unchanged (Hou *et al*., 2009, Sinko, 1999). Oral bioavailability is described as the amount of drug that reaches the systemic circulation unchanged after first pass metabolism (Kwon, 2002, Zhu *et al*., 2011). Therefore the main difference between the two is hepatic metabolism due to first pass effects of the liver (Figure 2.5). Oral bioavailability is a function of fraction absorbed (f_a), fraction escaping intestinal metabolism (f_g) and fraction escaping hepatic metabolism (f_h).

$$\text{Bioavailability (F)} = F_a \times F_g \times F_h \qquad \textbf{Eq. 2.2}$$

Intestinal absorption is measured in humans to give percentage human intestinal absorption (%HIA) or fraction absorbed (F_a), and oral bioavailability is measured by a variety of methods to give F as a percentage (Burton *et al*., 2002, Chiou, 2001).

It has been indicated that over half of compounds have the same absorption and bioavailability (Hou *et al*., 2009). Therefore, absorption prediction can potentially give an indication of bioavailability. This is advantageous due to difficulty in bioavailability prediction in drug discovery, because of the complexity of

metabolism and many other variables (Zhu *et al*., 2011, Metcalfe and Thomas, 2010).

2.7 Physiological and Physicochemical Factors Affecting Oral Absorption

It is important to distinguish what factors affect and how they affect oral absorption. This will aid in mechanistic understanding and improvement of resulting QSAR models that are built on the basis of previous knowledge of oral absorption. The three main areas affecting oral absorption are physicochemical, physiological and formulation factors (Figure 2.6). For the purpose of this thesis, only physicochemical and physiological factors affecting oral absorption will be discussed. Although formulation factors are of great interest in relation to oral absorption, they are not the focus of this thesis.

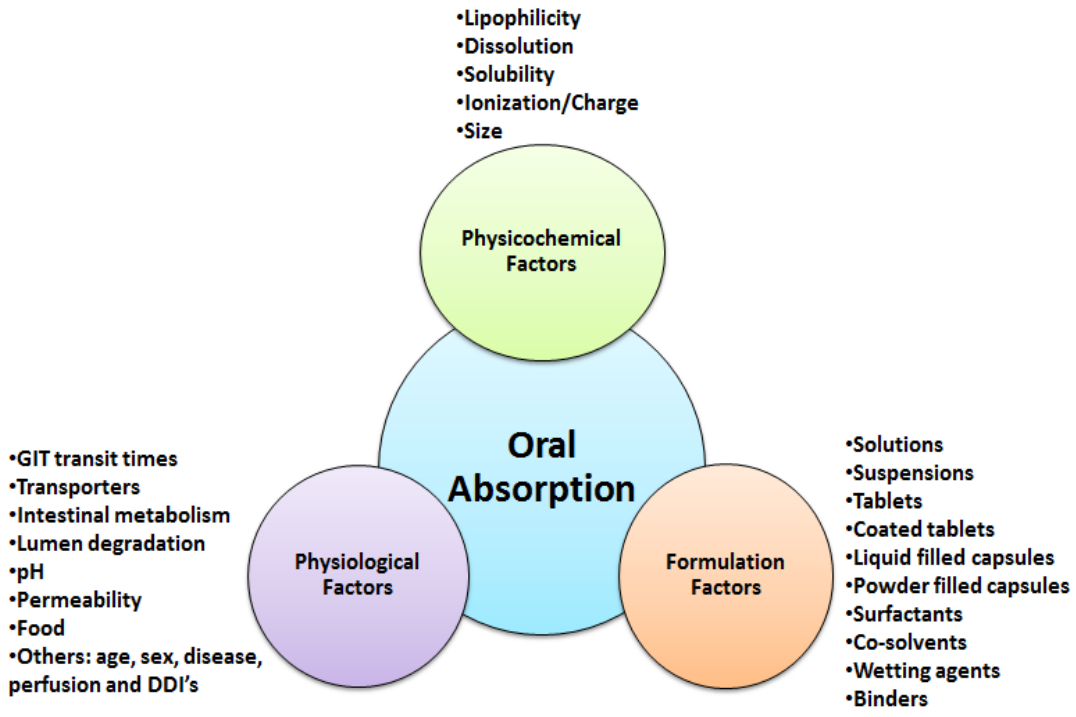

Figure 2.6. The main physicochemical, physiological and formulation factors affecting oral absorption adapted from (Ashford, 2007, Martinez and Amidon, 2002 and Darwich *et al*., 2010)

Physicochemical properties are drug characteristics that do not change, whereas physiological effects can depend on genetics or environmental factors of an individual. These factors can affect absorption independently or result in the change of other absorption-related factors directly or indirectly.

Physicochemical Factors Affecting Oral Absorption

2.7.1 Lipophilicity

Lipophilicity is one of the most important factors for intestinal absorption and determines other ADME properties of a drug (Testa *et al*., 2000, Arnott and Planey, 2012). In this context, lipophilicity determines the tendency of the drug molecule to partition either the lipophilic cell membrane or aqueous cell environment, and therefore can give an indication of membrane permeability and hence absorption. The partitioning of drug molecule between the two phases depends on intermolecular and intramolecular forces arising from the hydrophilic and hydrophobic functional groups on the molecule (Thomas *et al*., 2006, Testa *et al*., 2000).

The partition coefficient (logP) is the main measure of lipophilicity used in drug discovery (van de Waterbeemd and Gifford, 2003, van de Waterbeemd *et al*., 2001). It can be best described by the distribution of the drug molecule between two immiscible solvents, in particular 1-octanol and water. LogP can only be measured when the drug compound is completely neutral. However a majority of drugs are ionized at some point along the GIT due to the pH variations. Therefore logD is used and this is the logarithm of the apparent distribution coefficient between two immiscible solvents at a specific pH. LogD, although more complicated to measure, is usually the preferred property to calculate as it takes into account the ionization of the drug and pH (Ekins *et al*., 2000, Hou *et al*., 2007b).

2.7.2 Dissolution and Solubility

For a drug to be absorbed it needs to be dissolved in the GIT fluids. This depends on the dissolution rate and solubility of the compound. The dissolution rate is how fast the drug disintegrates and dissolves in the GIT fluids, whereas solubility is the concentration of drug that can be dissolved (saturated concentration) in a solvent at a specific temperature and pH. Therefore, the difference between the two is: solubility

is a thermodynamic process and is dependent on the physicochemical properties of the compound; whereas dissolution is a kinetic process, driven by solubility, and is a property of the compound in formulation also dependent on particle size and crystal state (Noyes and Whitney, 1897).

Solubility is a complex progression of energetically favourable processes such as crystal lattice energy, solvent cavity formation energy and solvation energy. Ultimately, the total energy depends on overcoming the intermolecular and intramolecular forces between drug molecules and between the solvent (GIT fluid) molecules, and formation of new intermolecular interactions between the drug molecule and GIT fluid, which results in the final solubility and availability for absorption (Lipinski, 2000, Wang and Hou, 2011).

Inadequate aqueous solubility can lead to poor, erratic, variable absorption. Therefore, solubility can be a rate limiting step for oral absorption. The impact of poor solubility is becoming more apparent as there are a growing number of new chemical entities (NCEs) that are practically insoluble in water (Lipinski, 2000, Savjani *et al*., 2012). The relationship between solubility and absorption-related parameters, such as lipophilicity, is normally negative (inversely related) (Buckley *et al*., 2012, Lipinski, 2000). Therefore physicochemical properties and some physiological properties such as GIT pH that are related to solubility are also related to oral absorption.

2.7.3 Ionization and Charge

Most drug compounds are weak acids or bases, so can exist as both ionized and unionized forms. The pH of the medium controls the proportions of ionized/unionized forms of the drug along the GIT and therefore absorption. The relationship between ionization and oral absorption is that unionized molecules are more permeable than ionized molecules and therefore the unionized state is the dominant form for passive diffusion (Brodie *et al*., 1957). The extent of ionization of a molecule at a specific pH can be measured using the acid dissociation constant, pk_a. At a specific pH, pk_a can be used to indicate the proportions of ionized and unionised forms of the molecule calculated from Henderson-Hasselbach equations, and is a function of the acidic and basic groups on the drug molecule. Originally it was

thought only the unionized neutral form of the drug could sufficiently cross cell membranes; nevertheless, it has been shown that some ionized drugs can still permeate cell membranes (Palm *et al*., 1999).

Some molecules have a permanent charge which is not influenced by pKa or pH. Examples of these permanently charged drug molecules are those containing quaternary ammonium groups, frequently found in neuromuscular-blocking drugs and antiviral agents. Due to the permanently positive charge these compounds have poor intestinal absorption but can potentially be absorbed via ion-pairing. Yet, there is speculation regarding the validity of this mechanism, as it can depend on physicochemical and physiological factors such as stomach content and also the strength of the ionic bond to the counter-ion (Miller *et al*., 2010, Jonkman and Hunt, 1983, Van Gelder *et al*., 1999).

2.7.4 Molecular Size

Molecular size is an important factor affecting intestinal absorption and biological activity. If molecular size increases, intestinal absorption decreases (Chan and Stewart, 1996). Molecular weight (MW) is the simplest indication of molecular size, although other parameters can be used such as surface area, molar volume and molar refractivity (Agatonovic-Kustrin *et al*., 2001). Molecular size hinders intestinal absorption of many drugs not just by the overall size and bulkiness of the molecule, but also by influencing other physicochemical properties such as lipophilicity and solubility. The importance of molecular weight is shown due to its inclusion in Lipinski's rule of 5, which states that poor absorption is likely if two or more of these conditions are satisfied: molecular weight >500, logP >5, number of H-bonding donor groups >5 or number of H-bonding acceptor groups >10 (Lipinski *et al*., 1997). There are exceptions to the molecular weight rule due to carrier mediated transport of larger compounds (Pang, 2003). Additionally, there are size restrictions for those compounds being absorbed via the paracellular route.

2.7.5 Gastrointestinal Transit Times and Food Effects

The small intestine transit time is usually constant (~3 hours) and not influenced by physical state, dosage form or food effects; the physicochemical effects of the drug influence the rate of absorption, not the small intestine transit time (Yuen, 2010, Davis *et al*., 1986). Gastric emptying, on the other hand, can be one of the main influencing factors for drug absorption for the majority of drugs (Prescott, 1974). It is defined as the time taken for the oral dose to move through the stomach and can be highly variable from 5 minutes to 2 hours (Ashford, 2007).

Gastric emptying depends on fed or fasted state (Ashford, 2007*)*, formulation, food composition (Charman *et al*., 1997), posture (Queckenberg and Fuhr, 2009), the drug itself (Fleisher *et al*., 1999) and the disease state of the individual (Heading *et al*., 1973). In particular, the absorption of some drugs, particularly acidic ones, can be directly correlated with the rate of gastric emptying (Prescott, 1974). With the knowledge of the effects of gastric emptying, some drugs are designed to take advantage of these aspects to increase their absorption.

2.7.6 Intestinal Gut Metabolism

Although the small intestine has mainly an absorption function, drugs can be susceptible to metabolism in the small intestine and this can contribute significantly to overall metabolism of the compound (Gertz *et al*., 2010, Thelen and Dressman, 2009). An overview of main metabolising enzymes in the small intestine is shown in Figure 2.7.

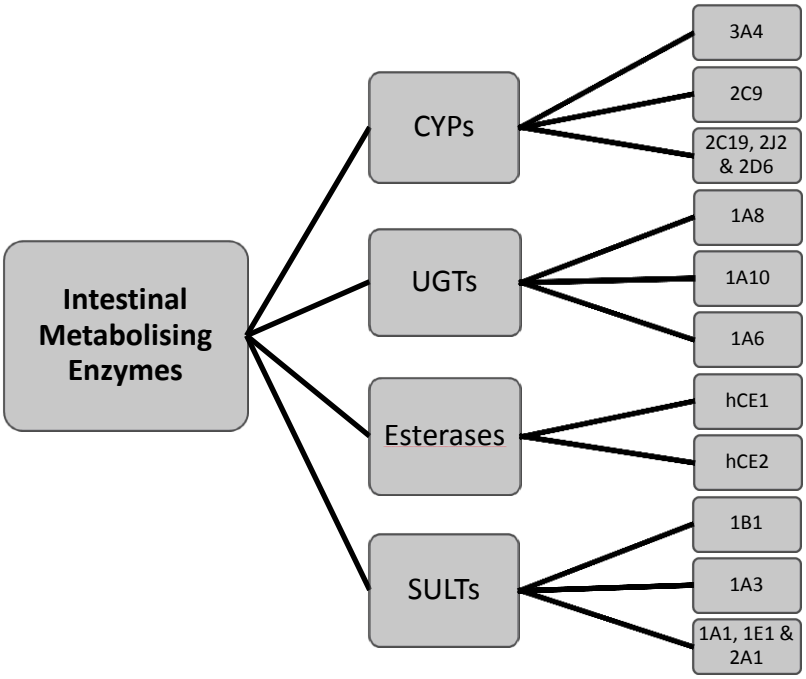

Figure 2.7. The main metabolising enzymes in the human GIT adapted from (Paine *et al*., 2006, Fisher *et al*., 2001 and Glatt *et al*., 2001)

Abbreviations: CYP, cytochrome P450, UGTs, Uridine 5'-diphosphate glucuronosyltransferases; SULTs, Sulfotransferases

Drugs can be degraded in the lumen by a variety of enzymes such as lipases, amylases, peptidases, as well as the enzymes from microflora present in the small intestine, all with overlapping or specific substrate specificity. The main metabolising enzyme present in the small intestine is the phase I cytochrome P450 (CYP) 3A4, making over 70% of all CYPs located in the small intestine with the highest abundance in the jejunum (Petri and Lennernäs, 2003, Thelen and Dressman, 2009).

The majority of enzymes in the small intestine are found in other places such as the liver. Some compounds, although highlighted as substrates for specific enzymes, are not susceptible to gut metabolism, but are metabolised by the same enzyme in the liver. Reasons for some compounds being susceptible to gut metabolism and others not, even though they are both enzyme substrates, could be the differences in biotransformation rate by the enzyme, solubility/dissolution rate, permeation rate, dose amount and substrate affinity (Fagerholm, 2007, Gertz *et al*., 2010, Lin *et al*.,

1999). The loss of the drug via gut metabolism is important to consider as it can result in the overestimation of absorption by *in silico* predictions.

2.7.7 Permeability

Permeability is the rate of absorption across the cell membrane in the small intestine. This is not the same as fraction absorbed, which measures the extent or amount of absorption (Burton *et al*., 2002).

Permeability is a popular measurement carried out in drug discovery using a wide variety of assays which will be discussed later on in this thesis. On the whole, there is a close correlation between the experimental permeability rate and overall oral absorption (Artursson and Karlsson, 1991, Lennernas, 1997). In general if permeability is poor then absorption and bioavailability are likely to be poor too, however there are exceptions.

Due to the close relationship between permeability and absorption, the same physicochemical factors affect both permeability and absorption such as lipophilicity, solubility and molecular size (Martinez and Amidon, 2002). Therefore, for passive transcellular absorption, permeability is higher for compounds that are lipophilic and unionized due to the lipophilic bilayer membrane. Additionally, physiological factors such as transporter abundance can greatly influence permeability and potentially affect the correlation with oral absorption.

As permeability is considered an important factor governing the absorption of many drugs, any experimental or computer model that can be utilised to indicate permeability can be used as a guide to indicate overall oral absorption.

2.7.8 Gastrointestinal pH

The pH along the GIT gradually increases from the stomach to the small intestine. Moreover, there are pH differences along the GIT in fed and fasted states (Table 2.1).

Table 2.1. Summary of the mean pH in fed and fasted states for the different areas of the gastrointestinal tract – adapted from (Dressman *et al*. 1998 and Horter *et al*. 2001)

GIT Region	pH range (Fasted)	pH range (Fed)
Stomach	1.4-2.1	3.0-7.0
Duodenum	4.9-6.4	5.1-5.2
Jejunum	4.4-6.6	5.2-6.2
Ileum	6.5-8.0	6.8-8.0

The main factors influencing pH are disease states, food, patient variability and the drug itself (Parsons, 1977, Martinez and Amidon, 2002). Similar to gastric emptying, these effects can be utilised and used to develop strategies to improve absorption, overcome degradation or improve ionization due to either fed or fasted states.

These pH differences in the GIT regions as highlighted in Table 2.1 can govern intestinal absorption due to different amount of ionization of the drug at the different pH's. There is a close relationship between pH and ionization, and this will govern intestinal absorption and many other properties such as solubility. Anything that affects the fraction of the unionized drug to be absorbed can decrease or increase the absorption. pH is a mitigating factor for the absorption of many drugs and important to consider in relation to oral absorption (Dressman, 1986, Charman *et al*., 1997).

2.7.9 Transporters

There are many influx and efflux transporters that can control membrane permeability of drug and endogenous compounds in the small intestine (Pang, 2003, Mizuno *et al*., 2003). The main transporter proteins present in the small intestine are shown in Figure 2.8. One of the main influential efflux transporters in the reduction of absorption and bioavailability of many compounds is a member of the ATP-binding cassette (ABC) transporters, permeability glycoprotein, P-gp. Compounds are frequently tested to see if they are substrates for the P-gp transporter due to its potential impact of lowering absorption and bioavailability of drugs (Chan *et al*., 2004, Zhang and Benet, 2001).

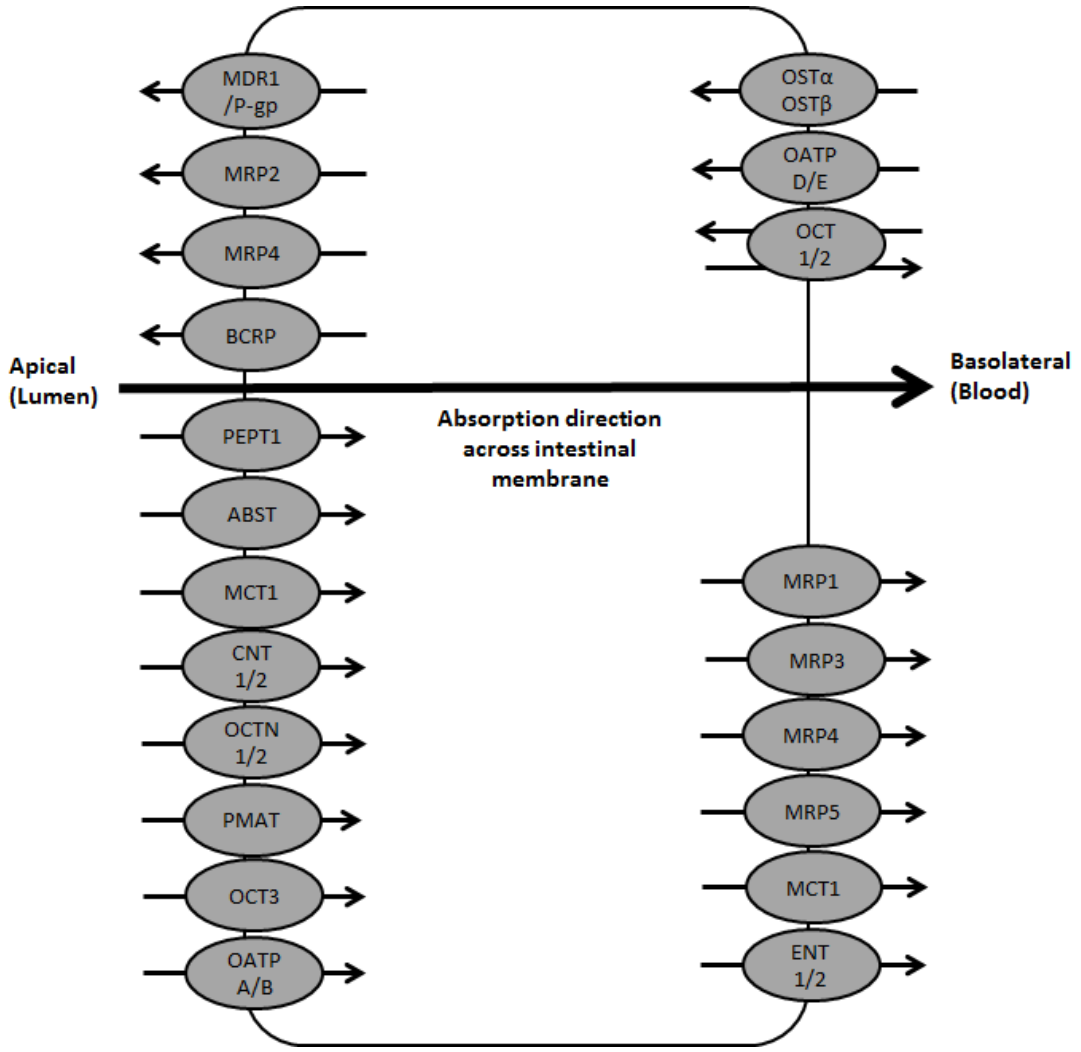

Figure 2.8. Diagram of major drug transporters expressed on the apical and basolateral intestinal membranes. Arrows denote transport direction.

Abbreviations; Multi-drug resistance protein (MDR)/P-Glycoprotein (P-gp); Multidrug resistance associated protein (MRP), Breast cancer resistance protein (BCRP), Monocarboxylate transporter protein (MCT), Peptide transport protein (PEPT), Organic anion transporting polypeptide (OATP), Organic cation transporter (OCT), Apical sodium-dependent bile acid transporter (ASBT), Concentrative nucleotide transporter (CNT), Electroneutral organic cation transporter (OCTN), Equilibrative nucleoside transporter (ENT), Organic solute transporter (OST), Plasma membrane monoamine transporter (PMAT). Adapted from (Custodio *et al*., 2008, Shugarts and Benet., 2009, Estudante *et al*., 2013; Morrissey *et al*., 2012, Sedykh *et al*., 2013, Yoshida *et al*., 2013).

The impact of transporters on absorption is difficult to predict. Factors such as specificity and substrate affinity can influence whether or not a drug will be a substrate. On the other hand, if a drug is a substrate for a transporter there are other

factors that can impact on absorption such as transporter abundance, contribution of transporter route to overall absorption, multiple substrate transport by other transporters, high substrate concentration and the kinetics (Giacomini *et al*., 2010, Sugano *et al*., 2010).

From a modelling perspective, transporter effects can give rise to incorrect predictions of passive oral absorption from molecular structure. With the growing research and increasing discoveries of transporter substrates it is important to attempt to take into account compounds that are transporter substrates in oral absorption models.

2.7.10 Other Factors Affecting Absorption

There are other factors that may be taken into account for the assessment of intestinal absorption; however these factors may not be essential for early drug absorption assessment in drug discovery. However, it does appear that there is a lack of research in some of these areas. A summary of these factors is presented in Table 2.2.

Table 2.2. Other factors that can affect oral absorption of drugs

Factor	How it can potentially affect absorption	References
Perfusion	Lack of sink conditions due to the slow removal of drug (i.e. perfusion) by the mesenteric capillaries.	Zhao *et al*., 2002
Exercise	Affects perfusion rate and appears to depend on type of exercise. Data in this area is limited.	Khazaeinia *et al*., 2000
Gender	Gastric emptying and pH have been shown to be different between genders. Pregnancy can increase hormone levels and physiological changes can affect perfusion.	Loebstein *et al*., 1997; Schwartz, 2003; Morris *et al*., 2003; Freire *et al*., 2011
Age	Higher pH in stomach and slower gastric emptying in new-borns. Slower intestinal transit, slower gastric emptying and postprandial pH response may differ significantly between geriatric population and younger healthier adults.	Koren, 1997; Gidal, 2006
Disease	Any disease or even surgery that affects the pH, mucosal enzymes or gastric emptying will affect absorption. For example, Crohn's disease decreases the surface area in small intestine and hence absorption.	Parsons, 1977; Gubbins and Bertch, 1991; Fleisher *et al*., 1999; Titus *et al*., 2013
Drug-Drug Interactions (DDIs)	Drugs that are absorbed via similar transporters will compete for that transporter, potentially resulting in altered absorption. This principle also applies to drugs metabolised by the same enzyme in the small intestine. DDIs can affect absorption indirectly by altering gastric emptying, gut motility and complexation with endogenous bile salts.	Richens, 1975; Welling, 1984; Fleisher *et al*., 1999

2.8 Experimental Assessment of Oral Absorption

In drug discovery there are many experimental assays that give an indication of oral absorption for compounds. The type of assay will vary in complexity, experimental throughput, cost and information gained from the experiment. The order of complexity increases from *in vitro*, *in situ* to *in vivo* assays. The assessment of oral absorption in drug discovery is usually a combination of these assays.

The experimental values from these assays can be used to build and validate computational models for the prediction of absorption for unknown compounds from structure alone. Computational models and simpler faster assays can be used to first screen a large number of compounds and, based on these results, fewer compounds are assessed by slower but more predictive assays (Balimane *et al*., 2000). As there are many assays, it must be noted that the value obtained must correlate with *in vivo* oral absorption in humans (Yu and Adedoyin, 2003).

An overview of *in vitro* and *in vivo* techniques where this data were utilised in the building of computational models in this thesis is described next.

2.8.1 In vitro Methods

There are a variety of *in vitro* techniques which measure the permeability of a drug using artificial membranes, cell monolayers and isolated intestinal tissues. The ideal permeability model for the small intestine mimics the physical and biochemical processes of intestinal absorption (Volpe, 2008). *In vitro* methods measuring permeability have been correlated with human intestinal absorption and these values can be used as surrogates in early drug discovery.

The main drawback of many *in vitro* assays is that they fail to take into account physiological and biochemical properties (Le Ferrec *et al*., 2001). Therefore, in some cases the relationship with the *in vivo* situation is difficult to establish. In spite of this, the isolation of one factor of oral absorption could be beneficial for problematic compounds. Although *in vitro* assays offer a fast and cheap alternative to the *in vivo* situation, the limitations of the methods must not be overlooked.

Apparent permeability (P_{app}) measured using cell monolayers is a popular *in vitro* method used in drug discovery and is usually measured in cm/s^{-1}. There are many different cell lines that can be used to measure permeability. Human colon adenocarcinoma (Caco-2) is a commonly used cell line (Balimane *et al*., 2000, Fogh and Trempe, 1975, Hidalgo *et al*., 1989), which displays biological and characteristic properties of the enterocytes of the small intestine such as the brush border and tight junctions (Artursson, 1990, Pinto *et al*., 1983, Hidalgo *et al*., 1989, Volpe, 2008). These cells can express a variety of transporters and metabolic enzymes, allowing other transport and metabolism mechanisms to be investigated (van Breemen and Li, 2005). Drawbacks of the Caco-2 cell line are inter-laboratory differences (which also holds for many other cell lines), variable transporter expression, long culture time, tighter junctions compared with *in vivo* situation and lack of mucus secreting goblet cells (Volpe, 2008, BriskeAnderson *et al*., 1997, Le Ferrec *et al*., 2001). Some of these problems have been resolved by other cell lines such as 2/4/A1, a rat intestinal epithelial cell line, which has leakier tight junctions (Matsson *et al*., 2005, Tavelin *et al*., 2003); also, the cell line HT29-MTX, a human colorectal adenocarcinoma cell line, is a co-culture of Caco-2 cells with mucus secreting goblet cells to study the effects of mucus on absorption (Hilgendorf *et al*., 2000).

Another cell line that has been gaining popularity is MDCK II (Madin-Darby Canine Kidney strain II) cells, due to shorter culture time (of 3-5 days), leakier tight junctions and low expression of transporters compared with Caco-2, making it an ideal cell line for passive permeability assessment even with species and tissue differences (Braun *et al*., 2000, Irvine *et al*., 1999, Avdeef and Tam, 2010, Varma *et al*., 2012). There are many similarities and differences between Caco-2 and MDCK cell lines. Despite this, there is a linear relationship between the permeability measured in the two cell lines, which has been shown using small compound sets (Irvine *et al*., 1999, Braun *et al*., 2000, Avdeef and Tam, 2010).

2.8.2 In vivo **Methods**

In vivo methods are studies of biological properties performed within living animals and humans. These methods integrate all the factors of absorption; therefore they offer the determination of intestinal absorption within a living system which other

methods lack. In spite of this, the integration of all components of absorption can be problematic as it is difficult to differentiate which one is the limiting factor (Lennernas, 2007, Le Ferrec *et al*., 2001). What is important to emphasise is that all the previous methods are meaningless unless they can characterize the fundamental principles of intestinal absorption as shown by *in vivo* studies.

Studies that measure the fraction absorbed are rare in humans due to the invasive nature of measurement. Therefore, the best way to determine intestinal absorption *in vivo* is using pharmacokinetic and/or mass balance studies in humans. Mass balance studies measure the amount of compound and/or drug-related material, usually radio labelled, excreted into either (or combined) urine, faeces and bile after oral administration. Additionally, the ratio of cumulative amount of drug excreted into urine after oral and intravenous administration can also be used (Varma *et al*., 2010, Zhao *et al*., 2001). Bioavailability studies can also be used as an indication of fraction absorbed, so long as the level of metabolism is negligible. The main problem with the *in vivo* assays used to determine fraction absorbed is that they are not suitable for high throughput screening (HTS) in drug discovery. Compounds with good drug like properties are tested *in vivo* in humans later in drug development and clinical studies (Gleeson *et al*., 2011).

3. Quantitative Structure-Activity Relationships (QSAR)

QSAR is the mathematical relationship between biological activity and structure. The structure of a chemical gives rise to different properties, and specific features are assumed to be responsible for the biological activity. The mathematical relationships between structure and activity can be established and then utilised to predict the activity of new chemicals. In other words, the biological activity of new chemicals can be predicted from structure alone without experimental measurement of activity (Figure 3.1). Moreover, this offers enormous advantages for the pharmaceutical industry to get an early idea of biological properties, such as oral absorption, before synthesis providing information for compound development.

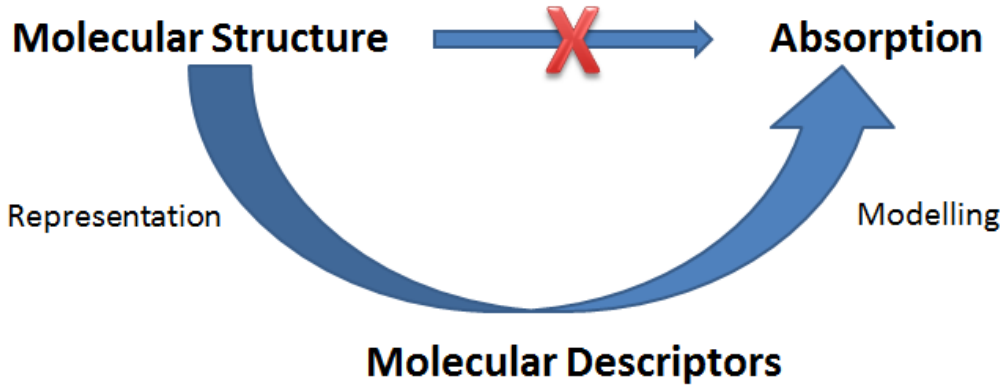

Figure 3.1. Graphical representation of the relationship between structure and activity for the underlying principles of QSAR

There are many studies that form the basis of modern QSAR today; the works by Hammett and Taft regarding the Hammett electronic constant and steric effects as well as many others made way for the work carried out by Hansch, who is considered the founder of modern day QSAR (Hammett, 1970, Hammett, 1935, Taft, 1952). The research carried out by Hansch and co-workers produced an equation that illustrated that the biological activity (Log 1/C) from chemicals could be described by a series of parameters: hydrophobicity and electronic effects (Hansch *et al*., 1962). From this work there have been vast developments and improvements for QSAR and these principles are the foundations for new approaches based on the machine learning and statistics.

The process of obtaining a suitable QSAR model can simply be put into five steps (Figure 3.2).

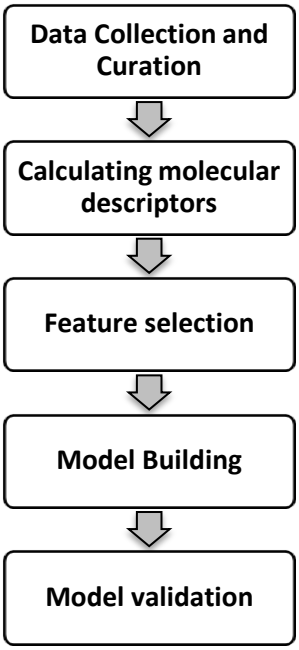

Figure 3.2. Steps to creating a QSAR model extracted from (Cherkasov *et al*., 2014)

3.1 Data Collection and Curation

The basis of QSAR models is the quality of data used to build them. Poor models result when data are of poor quality, therefore the collection of high quality data is essential. The quality of data can be a limiting factor in many QSAR models published in the literature (Egan and Lauri, 2002). In particular, oral absorption datasets are vulnerable due to the different methods to obtain the data plus other factors, such as formulation and solubility and variable absorption, which could also account for the differences.

Curation of the collected data is required to check and remove errors. In some cases published databases are used without checking primary references, therefore any potential errors can be copied from different studies in the literature (Hou *et al*., 2007c). Besides experimental variations, errors with the chemical structure can result in the calculation of incorrect molecular descriptors and can lead to significant reduction of predictability of models. Manual curation of structures can lead to

increased model predictivity (Young *et al*., 2008). Another common error in large datasets is the presence of duplicated compounds. Duplications can give a model artificial predictability and affect the model development by altering structural feature frequency. The work carried out by Tropsha highlights standardized steps for dataset curation (Tropsha, 2010, Fourches *et al*., 2010). Pre-processing initial data is essential to ensure a good starting point for model development and should result in higher predictivity of resulting models. This has been further emphasised by a recent publication by Cherkasov *et al* (2014). Additionally, recent work carried out by Golbraikh *et al* (2014) has developed a data modelability index (MODI), which estimates the feasibility of obtaining predictive QSAR models for binary data. This index utilises compound similarity and whether if similar pairs (calculated using Euclidean distances) of compounds have the same activity (Golbraikh *et al*., 2014).

3.2 Molecular Descriptors

Molecular descriptors are mathematical representations of the chemical structure used to derive relationships between structure and biological activity. The relationship between molecular descriptors and activity forms the basis of QSAR models. There are thousands of different molecular descriptors that can be used; these can vary from experimental measurements such as logP and solubility to purely theoretical ones based on quantum chemistry, graph theory, information theory and many more (Todeschini and Consonni, 2000). The choice of molecular descriptors can depend on several factors such as model interpretability, predictability and computational cost of molecular descriptor calculation. There are different types of molecular descriptors that have been categorised by dimensionality and presented in Table 3.1.

Table 3.1. Different types of molecular descriptors with examples – adapted from (Todeschini and Consonni, 2000)

Descriptor type	Descriptor	Examples
0D	Atom counts Bond counts *Simple counting of atoms/bonds*	Molecular weight Number of atoms e.g. H, C, Number and type of bonds e.g. H, C single/double bonds Number of ringed systems
1D	Structural fragments, fingerprints *Simple counting of fragments or atomic properties*	Number of carboxyl groups Number of hydrogen bond donors/acceptors
2D	Topological *Descriptors can be predicted using methods such as graph theory based on 2D structure*	Kappa shape indices Chi connectivity indices Topological polar surface area (TPSA) Weiner index BCUT/GCUT descriptors
3D	Geometric descriptors *Two types of descriptor based on internal or external orientation properties of structure, require 3D coordinates of molecule*	Potential energy descriptors e.g. solvation energy Size, shape and volume descriptors e.g. van der waals surface area, GETAWAY descriptors Polar Surface area
4D	Geometric flexibility descriptors *Descriptors derived from stereo-electronic or lattice representation, based on 3D but descriptors account for the different flexibility of same molecule i.e. different conformations*	Extension of 3D GRID or CoMFA methods, including Volsurf descriptors

A brief summary of the different types of molecular descriptors in relation to intestinal absorption is further described in the following sections. Rather than group 0D-4D molecular descriptors, in relation to oral absorption, for a better physicochemical understanding of the molecular descriptors, the descriptors have been grouped according to hydrophobic, electronic and steric effects, and a few examples relating to oral absorption have been discussed. Additionally this is a broad grouping, and there are many cases where molecular descriptors overlap.

3.2.1 Hydrophobic Descriptors

Molecular descriptors that relate to molecular interactions between polar and non-polar groups, including lipophilic partitioning, can be broadly termed hydrophobic descriptors.

LogP and logD are frequently used molecular descriptors that can be experimentally measured or theoretically obtained from the 2D structure. Although it gives an indication of the ability to penetrate a membrane, it is the lipophilicity parameter and other factors such as hydrogen bonding and ionization are also involved in defining the absorption. There is a positive contribution to the passive absorption of drug compounds with increasing lipophilicity (Zakeri-Milani *et al*., 2006). However, it has been shown that the relationship is non-linear with an optimum logP/D value for suitable absorption and solubility (Hansch *et al*., 1962).

There are other descriptors which are indirectly related to hydrophobicity. An example is polar surface area (PSA), which is a combination of steric (size) feature and the polarity of a molecule which is inversely related to hydrophobicity. PSA is the area of the van der Waals surface that arises from oxygen or nitrogen or to hydrogen atoms bound to these polar atoms in the molecule. Hence it is inversely related to hydrophobicity. This descriptor is often considered more suitable for the prediction of ADME than hydrophobicity descriptors such as calculated logP as it accounts for atoms that are shielded from other atoms within the molecule and internal hydrogen bonds (Palm *et al*., 1997, Clark, 1999). Another benefit is that this descriptor can be calculated from the molecules' 2D structure, so called topological polar surface area, so is much quicker to compute. Topological polar surface area has a high correlation with 3D PSA (Ertl *et al*., 2000).

3.2.2 Electronic Descriptors

Electronic descriptors give information relating the electronic distribution of a molecule. The electronic distribution or related electronic charge properties can be derived from empirical or molecular orbital calculations. Additionally electronic and steric indices can be combined to give charged partial surface area descriptors.

The main examples relating to intestinal absorption are dipole moments, polarizability and molecular orbital energies such as lowest unoccupied molecular orbital (LUMO) and highest unoccupied molecular orbital (HOMO) (Agatonovic-Kustrin *et al*., 2001, Norinder *et al*., 1999).

Dipole moments are measures of polarity plus other internal electronic effects of the molecules. Dipoles moments are created by atoms in the molecule with different partial charges that are a specified distance away from one another in various directions; it is a vector defined by the magnitude of the charge related to distance. The dipoles can vary in strength and type depending on the functional groups and their position on the molecule.

Polarizability descriptors give information on how susceptible the electron cloud of individual atom or molecule is to distortion by an external field. It implies that compounds that are electron rich with loosely bound electrons will have high polarizability. High polarizability is associated with high absorption (Norinder *et al*., 1999).

An example of electronic properties could be hydrogen bonding ability of a molecule. Counts of functional groups such as hydrogen bond donors and acceptors can give an indication of hydrogen bonding (Agatonovic-Kustrin *et al*., 2001). The influence of hydrogen bonding, which relate to inter and intra-molecular interactions between the compound and its environment, can dictate oral absorption of many compounds. In fact a high number of hydrogen bond acceptors and donors have a negative correlation with intestinal absorption and make up two of the rules of Lipinski's rule of five (Lipinski *et al*., 1997). More complex descriptors take into account hydrogen bonding strength, and internal hydrogen bonding can also be used (Abraham *et al*., 2002, Platts *et al*., 1999).

Finally, LUMO and HOMO energies are derived from quantum chemistry theories and relate to the reactivity of the compound. Although informative, they do require the 3D conformation of the molecule to be calculated. HOMO and LUMO energies relate to the electron donating and accepting ability of the molecule, respectively. A larger difference between these two energies results in a more stable absorbable compound (Agatonovic-Kustrin *et al*., 2001).

3.2.3 Steric and Topological Descriptors

Steric and topological descriptors give an indication of the size and shape. In particular, topological descriptors give an indication of size and shape through the connectivity of atoms in a compound. The size and shape can relate to a molecule's ability to pass through the enterocyte membrane or bind to a carrier transporter in the intestine.

Molecular weight is the simplest 0D size descriptor. A large molecular weight directly causes poor absorption as well as indirectly affecting other properties such as lipophilicity (Lipinski *et al*., 1997, Veber *et al*., 2002) .The shape of the molecule, which can be described by chi and kappa shape topological indices (Xue *et al*., 2004), can also influence absorption. These topological descriptors quantitatively encode information from molecular structural features. Chi or connectivity indices describe the number of atoms/fragments and branching of molecules including cyclic components of molecules (Hall and Kier, 1991). Kappa shape indices indicate the molecular shape in terms of cyclicity, branching and the position of the branches/cycles within the molecule, i.e. in the centre or closer to the extremities of the molecule. It is computed based on counts of one-bond, two-bond and three-bond fragments (Hall and Kier, 1991). Overall compounds that are planar and rigid with reduced molecular flexibility tend to have better absorption (Xue *et al*., 2004).

3.3 Selection of Molecular Descriptors: Feature Selection

As seen from the previous section, there are thousands of ways to characterise a chemical compound computationally to produce numerous molecular descriptors. In order to develop a model that is robust and has a high predictive power, the selection of molecular descriptors is very important. By considering the process of oral absorption, any descriptors that can influence this property in any way will be useful in creating a predictive model (Jensen *et al*., 2005). Identifying the relevant descriptors correlating with intestinal absorption can be carried out using statistical feature selection methods, or additionally, educated assumptions can be made about the process of oral absorption and the physiological and physicochemical factors that influence it, in order to choose the useful descriptors (Suenderhauf *et al*., 2011).

Feature selection is frequently used in QSAR to selectively reduce the number of molecular descriptors (independent variables) used to accurately describe the property of interest (dependent variable) (Wong and Burkowski, 2011, Xue *et al*., 2004). Feature selection is important for several reasons. Firstly, fewer molecular descriptors increases interpretability and understanding of resulting models (Ghafourian and Cronin, 2005, Liu, 2004). Secondly, feature selection can provide improved model performance for the prediction of new compounds (Dudek *et al*., 2006, Xue *et al*., 2004). Finally, feature selection can reduce the risk of overfitting from noisy redundant molecular descriptors (Goodarzi *et al*., 2012).

Feature selection can be split into two broad categories; data pre-processing or embedded methods. Data pre-processing feature selection involves reduction of molecular descriptors before model building and can be further split into filter and wrapper techniques, whereas embedded methods incorporate the feature selection into the training and building of the model (Goodarzi *et al*., 2012, Saeys *et al*., 2007). Filter techniques usually involve calculating a relative score of the molecular descriptors and ranking them in order of best score, and the descriptors that are at the top of the list are then used as input for model building. Filter methods offer a fast and simple way to select important descriptors. In addition, because they are independent of the algorithm, the score for each descriptor only needs to be calculated once, and the selected descriptors can be used as input for a variety of algorithms. A disadvantage of univariate filter methods is they fail to account for interactions between independent variables, as most measure the correlation between the dependent variable and each independent variable separately. This can be overcome by multivariate filter methods which take into account independent variable interactions (Saeys *et al*., 2007).

Wrapper techniques consider a number of candidate subsets of molecular descriptors, evaluate each of these based on the predictive performance of a model built from that descriptor subset, and eventually select the descriptor subset with the best predictive performance (Kohavi and John, 1997). Wrapper techniques are usually much more computationally expensive than filter techniques, but unlike many univariate filter techniques, they take into account independent variable

interactions (Goodarzi *et al*., 2012, Saeys *et al*., 2007). There are now methods that combine filter and wrapper techniques to create successful hybrid feature selection techniques (Wegner *et al*., 2004).

3.4 Development of QSAR Models

To develop a QSAR model, the purpose and intended use of the model need to be established. This will help in the selection of the appropriate method based on the level of interpretability and predictability taking into account cost-effectiveness. QSAR models can be mainly split into two categories; unsupervised and supervised methods.

Unsupervised methods do not distinguish between dependent and independent variables (i.e. there is no special dependent variable to be predicted); rather, their goal is to discover patterns and associations between the variables to potentially define possible groups. These groups and patterns may be used later in order to make predictions using supervised methods. The unsupervised methods are applied under the assumption that similar compounds (with similar descriptors) will have similar activities, and so will be grouped together and reveal meaningful patterns not seen in the raw data. Unsupervised methods, such as clustering methods, can also be used as pre-processing techniques to select training and validation sets (Martin *et al*., 2012, Golbraikh and Tropsha, 2002b). Another example of unsupervised methods is principal component analysis (PCA). PCA is used to produce new independent variables (principal components), which then can be used as molecular descriptors in model development by supervised methods (Lauria *et al*., 2009). Some types of data mining methods are broad enough to include both supervised and unsupervised algorithms, such as neural networks (Gini *et al*., 2004).

With supervised methods, the dependent variable is known and it is this information plus the input from the independent variables (molecular descriptors) that guide the supervised method to predict the dependent variable. Supervised methods are more common in the QSAR field and for ADMET prediction. These methods can be further split into regression or classification methods, where the dependent variable to be predicted is numerical (continuous) or categorical (e.g. ‘high’ or ‘low’), respectively (Dudek *et al*., 2006). There are factors to consider when choosing

between regression or classification methods, such as data quality, interpretability and purpose of the intended outcome. For example, a regression model based on data with high variability will produce a poor predictive model, whereas classification into two classes can remove noise and variability and could produce a better model, but some information can be lost. As another example, categorical classification methods would probably not be suitable for a chemical series of lead compounds which only differ by a functional group, as a classification model would be too broad, and therefore a regression model to distinguish between very similar compounds predicting a precise numeric value would be more suitable.

A brief summary of the main unsupervised and supervised methods used in QSAR is presented in Figure 3.3 and the following text:

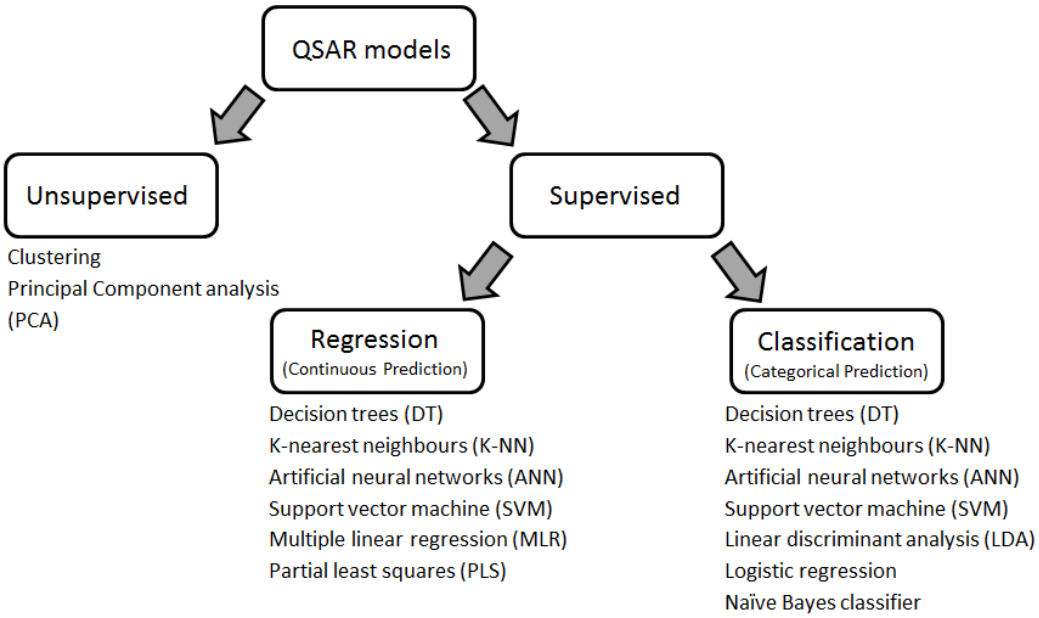

Figure 3.3. Summary of typical methods used for QSAR models adapted from (Dudek *et al*., 2006)

It must be noted that several types of methods mentioned in Figure 3.3 can be utilised for both regression and classification analysis. A good example is decision trees, particularly the classification and regression trees (C&RT) algorithm. This method is able to predict either the numerical value or the (categorical) class of the

compound (Breiman *et al*., 1984). Supervised methods can either be linear or non-linear, and this is something else to consider when selecting a model. Linear models predict the dependent variable as a linear function of the molecular descriptors. A good example is multiple linear regression. Linear methods can offer high accuracy and interpretability for models. However, some molecular descriptors are not linear in relation to the dependent variable, which can restrict the predictive accuracy of the resulting linear models. For example, logP and logD have non-linear relationships with intestinal absorption. Therefore, non-linear methods can be used to take into account the non- linear function of the molecular descriptors. However, non- linear methods such as SVM and ANN can be harder to interpret and prone to over fitting (Dudek *et al*., 2006).

3.5 Validation of QSAR Models

Once a model has been built using the training set, it needs some sort of validation. Model validation has received much attention for the acceptance of QSAR models (Gramatica, 2007), particularly from a regulatory point of view. For consideration of the use of QSAR models for regulatory purposes, the OECD principles were established to offer international agreement or a benchmark of model validation (OECD, 2007). In particular, model validation requires appropriate measures of goodness-of-fit, robustness and predictivity. The training sets are assessed on goodness-of-fit and robustness and the internal and external validation set assesses the predictivity of the model.

Performance of models can be quantified in some form of accuracy measure. There are many accuracy measures utilised in QSAR. For regression, the statistical parameter most popular is r^2, particularly used in MLR, which assesses the rectilinear relationship between the dependent and independent variables. The Fisher (F) statistic is used to assess the statistical significance of the regression model. As the F-value becomes higher, the greater the probability is that the equation is significant (OECD, 2007). The error of the regression methods can be assessed based on the difference between the model's predicted values and the observed values. Parameters such as root mean square error (RMSE), Mean absolute error (MAE) and mean fold error (MFE) are just a few examples of measurements used in QSAR

studies to measure performance of the training and validation sets. The lower the error the better, as there is less difference between the observed and predicted values (Aptula *et al*., 2005).

For the assessment of classification models, originally utilised for medical testing, the Cooper statistics (Cooper *et al*., 1979) are commonly used for the classification of QSAR models (Hou *et al*., 2007c, Hou *et al*., 2007a). This is based on percentages of the overall correct predictive performance in relation to the number of compounds in the training or validation set, or can be further split to calculate parameters that give values for predictive accuracies associated with different classes in the classification model, called specificity and sensitivity (Baldi *et al*., 2000). Other parameters used in QSAR validation are Matthews correlation coefficient (Matthews, 1975), Youden J statistic (Youden, 1950) and kappa weighed index (Cohen, 1968).

The measures mentioned above can be used to calculate the accuracy of a model for training set (model fit) and its predictive performance on validation sets. The performance of the models' predictive accuracy the training set is likely to be over optimistic of the model's true predictive power. Therefore the predictive power should be assessed using internal validation of the training set and an independent external validation set. An internal validation or sometimes called cross-validation are common statistical techniques, where different proportions of the compounds in the training set are iteratively removed from the training set and predicted using the developed model. Examples of these methods are leave one out (LOO), leave more out (LMO) or bootstrapping (Gramatica, 2007). There has been much criticism of solely using cross validation values to assess model predictability. In fact, these methods tend to overestimate the predictive ability of the resulting model as well as being computationally expensive. However, others have argued that when the sample size is small, keeping back an external validation set of compounds is wasteful (Hawkins *et al*., 2003, Hawkins, 2004). It has been shown that a high value of internal validation (Q^2) using LOO does not necessarily correlate with high predictive power (Golbraikh and Tropsha, 2002a, Golbraikh and Tropsha, 2002b). Additionally Y-scrambling or Y-randomization can be used to highlight if the model

occurred by chance correlation due to redundant molecular descriptors (Tropsha *et al*., 2003).

An external validation set is a set that is independent of the training set and is used to determine whether or not the model built has a good predictive performance for a new set of compounds. If there is a correlation between the predicted and observed values for the external validation set, then over fitting of the model has not occurred. Using an external validation set is considered by some authors the best way to evaluate the predictive accuracy of a model (Golbraikh and Tropsha, 2002a, Golbraikh and Tropsha, 2002b). However, this does not mean the internal validation techniques are not useful, they are vital for model development; but when they are used with an external validation set a more complete assessment of model validation can be achieved (Gramatica, 2007). What is apparent is that some form of model validation is required and justified. For methods that require external model validation, firstly the dataset needs to be split into a training set and validation/test set. The training set is used for building and optimising the model and the validation set is used as a validation of the predictive power of model using a dataset not used for model development. The splitting of the dataset into training and validation sets can be random or produced by using algorithms such as the one proposed by Kennard and Stone (Kennard and Stone, 1969). It is important to make sure that compounds in the training set are structurally diverse and cover a large chemical space, and the validation set compounds must be similar to the ones in the training set to avoid extrapolation (Martin *et al*., 2012).

In addition, the applicability domain (AD) estimation can be calculated. Broadly speaking, the applicability domain is essentially used to assess the reliability of a model for future predictions by defining the chemical coverage of the training set compounds, in comparison to those compounds being predicted by the model (Netzeva *et al*., 2005). In other words, according to the similarity approach for AD, prediction set compounds must be reasonably similar to those in the training set; otherwise extrapolation will occur and a prediction with less confidence is expected. The definition of chemical similarity can be subjective; however, in order to see if the model can interpolate for the validation set compounds, a variety of techniques

such as PCA analysis, probability density and distance-based methods can be carried out (Jaworska *et al*., 2005, Sahigara *et al*., 2012). The applicability domain is calculated in model validation but can transcend into model development, and can be calculated constantly to ensure any new chemicals predicted using the proposed model fall within the scope of the applicability domain of the model.

4 In silico Models for the Prediction of Oral Absorption

There have been many publications in the literature that focus on the prediction of oral absorption. These can be subdivided into categories and the use of each type of model will depend on many factors. The main factors that can determine which category of model is used are the availability of data and the level of predictability and interpretability needed at any particular stage in drug discovery. As there are many examples of oral absorption models in the literature, only those relevant to this thesis and the problems in hand were selected; but it is appreciated there are many more examples that could be used. The recent publications on the prediction of oral absorption highlight that this area is an important research topic for the pharmaceutical industry.

4.1 Oral Absorption Models are Built Using Highly Unbalanced Datasets

Early oral absorption models were based on small datasets of usually fewer than 100 compounds (Wessel *et al*., 1998, Niwa, 2003, Norinder *et al*., 1999). Small datasets are not ideal, as models built using them are affected by lack of generalizability for new compounds. Although more recently larger datasets have been published to potentially overcome this lack of generalizability, in general datasets still share the same problem (Zhao *et al*., 2001, Hou *et al*., 2007c, Cao *et al*., 2012). The problem is that datasets contain many more highly-absorbed compounds compared to poorly-absorbed compounds, creating a highly unbalanced dataset.

In general, oral absorption datasets published in the literature contain around 80% of highly-absorbed compounds (Wessel *et al*., 1998, Hou *et al*., 2007c, Zhao *et al*., 2001). In this case, highly-absorbed means any compound with 50% or more fraction absorbed. The main reason for this is that larger numbers of highly-absorbed compounds are amongst the marketed drugs that constitute the datasets (Wessel *et al*., 1998, Zhao *et al*., 2001). Furthermore, the vast majority of percentage oral absorption data are obtained from clinical trials, where it is expected for compounds to have good absorption in order to have reached this stage. Additionally, the lack of

published data representing poorly-moderately absorbed compounds is thought to contribute to this unbalanced dataset problem as well (Gleeson *et al*., 2011).

Due to the high number of highly-absorbed compounds compared with poorly-absorbed compounds, models produced will have better predictability for the majority highly-absorbed compounds and will be poorly predictive for the poorly-absorbed compounds. For many models in the literature this statement is true. Although this imbalance might be initially appealing for model building, it is not suitable for numerous reasons.

Firstly, models should be able to distinguish equally between highly and poorly-absorbed compounds and not be influenced by the dataset distribution. A model is not useful to predict high or low absorption if the majority of cases are highly-absorbed, as it will be unable to predict well those that are poorly-absorbed due to their under-representation. Secondly, these models in the literature based on these biased, unbalanced-class datasets are not representative of a true industry scenario at present, where there are more drug candidates with poor absorption. There are more compounds that are poorly-absorbed due to the increasing number of larger, more lipophilic and poorer solubility compounds being designed in drug discovery (Lipinski, 2000, Leeson and Springthorpe, 2007).

This pattern of dataset imbalance is seen throughout many datasets in the literature and not just oral absorption datasets (Czodrowski, 2013, Eitrich *et al*., 2006). Furthermore, for oral absorption models in the literature, dataset imbalance is stated as a problem; however, there are few studies that attempt to resolve it (Niwa, 2003, Yan *et al*., 2008). In spite of this, there are exceptions where a resolution to overcome the unbalanced class distribution of datasets was considered; however, there is no extensive comparison of models produced when using unbalanced and balanced datasets for oral absorption prediction (Talevi *et al*., 2011, Hou *et al*., 2007a, Bai *et al*., 2004).

4.2 Oral Absorption Models are Based on Passively Absorbed Compounds with Good Solubility

The compounds used to build absorption models cover the different absorption processes from passive to active transport of chemical compounds. They therefore will give rise to models with a large applicability domain. In spite of this, compound data are sometimes removed in order to help produce useful models for specific predictions, in the case of oral absorption models for the prediction of passive absorption (Hou *et al*., 2007a, Hou *et al*., 2007b).

The main oral absorption models in the literature remove compounds that are absorbed via carrier mediated transporters, have solubility issues, low data reliability and software's inability to calculate molecular descriptors (Zhao *et al*., 2001, Hou *et al*., 2007c). According to the works by Tropsha and co-workers, data curation is one of the essential steps in model building to ensure a homogenous dataset is produced suitable for modelling (Fourches *et al*., 2010, Tropsha, 2010). Although an essential step in data curation, the removal of these 'outlier' compounds as specified above could result in oral absorption models with reduced applicability and generalization to new compound sets. Additionally, there are other implications which are discussed next.

Firstly, a particular example of data curation is used in the work by Hou and co-workers. For modelling, 95 compounds were excluded from a training dataset of 647 compounds. These compounds were identified to undergo absorption mechanisms other than transcellular, had solubility problems, missing logD values or had a permanent quaternary ion in the structure. Their justification on compound removal was to guarantee the accuracy for passive absorption models. Although the removal of these 'outliers' has resulted in a homogenous dataset with compounds undergoing similar biological processes, there are potential issues. Firstly, they removed paracellular compounds, also defined as passively absorbed. Secondly, with recent research on transporters, there are more compounds being identified as undergoing transport mechanisms. Finally, those compounds with a permanent quaternary ammonium ion were initially excluded but then added to the training as an additional rule for compounds prediction, therefore aiding the statistics of their model. In spite

of this, many other QSAR works have utilised the dataset with compounds exclusions (Yan *et al*., 2008, Shen *et al*., 2010).

The main question is whether or not the removal of these compounds is necessary. Although Hou stated compound removal was carried out to guarantee the accuracy for passive absorption models, there was no comparison with and without these outliers. The comparison of models with and without the compounds absorbed via carrier mediated transport was carried out by Taveli *et al*., 2011. The removal of these compounds did not result in a significant difference between r^2 values for the training set of models built, where r^2 of 0.659 and 0.663 were obtained using the full dataset and the dataset after removal of outliers, respectively. This study suggests that non-passively absorbed compounds could be included in models to increase model applicability.

Additionally, it has been hypothesised that compound absorption can involve the co-existence of different absorption mechanisms, but oral absorption may be governed by the dominant process (Sugano *et al*., 2010, Smith *et al*., 2014). In fact, Reynolds identified some highly-absorbed compounds that were also substrate for efflux transporters and poorly-absorbed compounds identified as substrate for influx transporters (Reynolds *et al*., 2009). With these studies in mind, plus the increasing research into carrier mediated absorption, the inclusion of these compounds would be attractive for industry purposes.

Compounds with solubility and dissolution problems are also commonly removed or not included in oral absorption models (Egan *et al*., 2000, Reynolds *et al*., 2009). This is because inadequate aqueous solubility can result in poor and variable absorption, making absorption prediction of these compounds more difficult and with higher errors (Zhao *et al*., 2001). However the simplification of resulting models will reduce the applicability and may impact the potential generalizability of the resulting models.

4.3 Oral Absorption Models Can Be Built From a Selection of Thousands of Different Molecular Descriptors

There are many molecular descriptors that can describe chemical structures available from various software programmes. Therefore, the selection of the most important and relevant molecular descriptors is vital. The reduction of the set of molecular descriptors increases interpretability and simplicity of resulting models, as well as improves model performance and reduces risk of over-fitting (Goodarzi *et al*., 2012).

The majority of oral absorption models in the literature utilise some form of feature selection method, either in pre-processing or embedded manner. There is a variety of research in the literature, and this varies greatly with the focus of feature selection. There are fewer studies in the literature that compare different feature selection techniques and compare the molecular descriptors chosen by the different techniques for oral absorption (Wegner *et al*., 2004, Suenderhauf *et al*., 2011). The majority of studies, however, focus on obtaining a model with high predictive accuracy, and do not have feature selection as their primary focus (Hou *et al*., 2007c, Talevi *et al*., 2011).

Examination of the wide range of feature selection methods utilised in individual oral absorption models reveals that different feature selection methods should be tried and evaluated for the dataset at hand. For many oral absorption models only one feature selection method is applied to each study; these vary from simple filter methods through to those that have embedded feature selection in model training. Common examples of the latter type of feature selection approach, where studies focus on high predictive accuracy, used in the literature are Genetic algorithm (GA) combined with neural networks (Wessel *et al*., 1998), support vector machine (Yan *et al*., 2008) and multivariate adaptive regression splines (MARS) (Hou *et al*., 2007c).

In contrast to this, Xue and co-workers considered three different datasets, including one involving the prediction of oral absorption. They used recursive feature elimination (RFE) for feature selection and Support Vector Machine (SVM) to classify compounds. They compared the results with and without the feature

selection method and found that, for oral absorption, improved accuracy was obtained when the feature selection method was used (Xue *et al*., 2004). Whereas the study by Xue *et al*. only used SVM alongside RFE feature selection, Suenderhauf *et al*. used a variety of modelling techniques including C&RT, SVM, and chi-squared automatic interactor detector (CHAID). These modelling methods were used alongside a variety of feature selection methods such as best first feature selection (BFS), a greedy hill-climbing algorithm, linear correlation analysis and decision tree splitting criteria (Suenderhauf *et al*., 2011). Suenderhauf *et al*. found that feature selection before model training did not improve model accuracy, in contrast to the findings of Xue *et al*. A possible conclusion is that in Suenderfauf *et al*. the preprocessing feature selection was less effective because two of the classification algorithms (C&RT, CHAID) perform embedded feature selection, reducing the need for pre-processing feature selection.

4.4 Oral Absorption Models are a Balance Between Interpretability and Predictivity

There are a variety of computational methods that have been utilised for the prediction of oral absorption. These vary from simple rules of thumb to more complex methods such as support vector machine and artificial neural networks. The choice of method can depend on many factors such as the dataset and computational time, but most importantly interpretability and predictability. A balance between interpretability and predictability is required in order to obtain a valuable model for the prediction of intestinal absorption that is also user friendly for application.

The simplest models are those based on rules of thumb. These consist of analysing large compound datasets and finding patterns of properties or specific structures that correlate with good oral absorption. The main advantage of this type of model is the simplicity and interpretability for the non-specialist; and this approach is sufficient for high throughput screening (HTS) for a quick approximation, which can highlight potential unsuitable/suitable compounds for further drug development. The main example is Lipinski's Rule of five (Lipinski *et al*., 1997). Originally used for assessing drug-likeness, the rules were applied for oral absorption. Lipinski stated that poor absorption is highly likely to occur if two or more of the following rules

were satisfied: if molecular weight >500 Da, sum of OH and NH hydrogen bond donors >5, calculated logP (C LogP) >5 and sum of N and O atoms as hydrogen bond acceptors >10.

Even with the simplistic nature of the rule of five, there have been criticisms stating these rules are too simple to predict intestinal absorption, due to its complexity. In addition, the rules can lead to a high number of false positives i.e. those that satisfy the rules indicating they should be poorly-absorbed but are actually highly-absorbed due to carrier-mediated influx transporters. Therefore, the rule of five is only suitable for passively absorbed compounds (Yalkowsky *et al*., 2006, Lipinski *et al*., 1997). In addition, other works in the literature have indicated that some properties not included in the rule of five, such as number of rotatable bonds, PSA (polar surface area) and solubility should be considered too, as all are important for good absorption (Lobell *et al*., 2006, Clark, 1999, Palm *et al*., 1997, van de Waterbeemd and Kansy 1992).

Although a qualitative prediction for oral absorption is quick and simple, a quantitative approach allows a deeper mechanistic understanding of the different processes of oral absorption. In general, earlier oral absorption models appear to favour prediction of a numerical/continuous value, rather than classification into high or low absorption groups (Talevi *et al*., 2011, Hou *et al*., 2006). Numerical (regression based) models of oral absorption have produced acceptable predictive accuracy and remain popular. In spite of this, it has been argued that, due to the variability of the data, the numerical prediction of HIA may not be suitable, with experimental errors of HIA% being as high as 20% for some compounds (Klopman *et al*., 2002). This could really affect numerical predictions, unlike binary classification, which could cover the error of the data as long as the variation does not cover the threshold that distinguishes between the two classes.

Early studies of Wessel (1998) and Zhao (2001) have used methods such as MLR and ANN for numerical HIA prediction, and these methods are used frequently by other studies (Niwa, 2003, Talevi *et al*., 2011). ANN is a good technique to use for modelling parameters with complex relationships such as oral absorption. However, it can prone to over-fitting, plus it is difficult to interpret as there is no QSAR

equation produced from the output of the algorithm. Other methods such as PLS, SVM and Multivariate Adaptive Regression Splines (MARS) have also been adopted for the numerical prediction of oral absorption (Deconinck *et al*., 2007, Yan *et al*., 2008, Norinder *et al*., 1999).

In general there have been fewer classification models for oral absorption in earlier literature studies, however this has been growing over the past decades (Hou *et al*., 2006). The use of classification methods such as SVM has appeared in publications for the prediction of oral absorption (Shen *et al*., 2010, Hou *et al*., 2007a). One of the highest predictive accuracy models for oral absorption classification was obtained using SVM. Hou *et al* (2007a) achieved 98% classification accuracy for the validation set (n=98) using SVM developed using a training set of 480 compounds. Although SVM and ANN methods achieve high predictive accuracy, the models built by these methods are difficult to interpret, and hence they are labelled 'black box' techniques (Tian *et al*., 2011). Other methods for categorical prediction include DT (decision trees). These methods offer a balance between interpretability and predictability, by showing a simple visual representation of the model as well as having good predictive accuracy (Deconinck *et al*., 2005, Hou *et al*., 2007c).

In addition to the type of prediction to be made there is another factor to be considered. Methods can either be linear or non-linear. In general, the use of linear methods, in particular MLR was popular in earlier oral absorption models; in fact Wessel *et al* (1998) highlighted the comparison between linear vs. non- linear methods by modelling oral absorption using MLR and ANN. The models achieved a RMSE of 35% and 16% for validation set for linear regression and neural networks respectively. Therefore, this indicates that non- linear methods can outperform linear methods. This result has also been shown in the majority of studies of oral absorption models (Wessel *et al*., 1998, Talevi *et al*., 2011, Yan *et al*., 2008). Non-linear methods have been shown to work better due to the larger overestimation of poorly-absorbed compounds by the linear methods compared to non- linear methods. In addition, non-linear methods take into account the non-linearity of some molecular descriptors such as logP/D and ionization (Reynolds *et al*., 2009). Despite this, Zhao and co-workers found that, although there were lower predictive errors for

non-linear models as expected, the regression coefficients between linear and non-linear models were similar (Zhao *et al*., 2002).

No model is perfect; therefore models will have incorrect predictions. These can either be false negatives or false positives. In the context of oral absorption, false negatives occur when classfication models predict an observed highly-absorbed compound as poorly-absorbed and false positives occur when models predict an actually poorly-absorbed compound as highly-absorbed. Therefore, in addition to considering interpretability and predictability of the models produced, the emphasis on which type of error is more important to reduce in drug discovery is under debate, as both are detrimental.

False negative predictions give rise to missed opportunities of potential new blockbuster drugs, and on the other hand false positives can give rise to expensive unsuitable compounds. There appears to be a lot of business emphasis on reducing the number of false negatives in drug discovery due to the potential of missing the next potential drug and therefore potential loss of revenue (Malo *et al*., 2006). A reduction in the number of false negatives is favoured in most publications, as in practice they are more difficult to assess and highlight, so a model with as low as possible false negative rate is preferred (Zhang *et al*., 2000).

However, despite this, reducing the number of false positives could be considered equally as important or more important for cost-effectiveness reasons. If a drug is misclassified as highly-absorbed when in fact it is poorly-absorbed (false positive), more time, effort and money are invested to investigate and reveal the compound's true class with further tests. Although there are few publications indicating that false positives need to be decreased rather than the business need of reducing false negatives, with the spiralling cost of drug discovery, a future consideration for many companies may be to reduce false positives and therefore to become more cost and time effective (Cummings, 2006, Oprea, 2000). A suitable balance of errors, depending on the context of the problem and the intended outcomes may be the answer to reduce time and money testing unsuitable drugs, compared with reducing the potential for missed opportunities of new drug candidates, as long as there are

still a high number of true positives being discovered (White, 2000, Rydzewski, 2008).

Overall, from the literature, the justification of a modelling method can be based on the successful use of the method for modelling oral absorption and/or the consideration of the balance between interpretability and predictability of the resulting model.

4.5 The Relationship Between Oral Absorption and *in vitro* Permeability is Determined Subjectively

From a modelling perspective, the prediction of fraction absorbed in humans would be the best approach rather than prediction of *in vitro* measures of permeability. However as stated previously, studies in humans are carried out later, on fewer compounds, therefore this causes a potential problem for oral absorption datasets. In terms of future developments, it is expected that only a small number of data will be added for the validation and model improvement over time (Egan *et al*., 2000). Plus the chances of the new additions being highly-absorbed are high, therefore this adds to the data imbalance problem discussed earlier.

Permeability and related parameters are now frequently experimentally measured in drug discovery, in particular using Caco-2 and MDCK cell lines (Irvine *et al*., 1999, Volpe, 2008). As discussed previously, the correlation between the results of these cellular assays and human absorption allows an indication of human absorption earlier on in drug discovery. Due to the increase in HTS, there are more permeability data that could be potentially added to the dataset to possibly create an even distribution of data.

The relationship between the permeability and fraction absorbed of a drug in humans can be determined numerically (regression) or categorically (classification). From a classification perspective, a permeability threshold can give an indication of high or low intestinal absorption (absorption class). The permeability thresholds defined in the literature vary greatly (Table 4.1) and the majority of studies appear to set the permeability threshold subjectively from a visual inspection of the graphical fit, rather than using an objective method using analysis of existing datasets.

Table 4.1. Examples of permeability thresholds determined by the literature

Study	Cell line	Papp threshold (x 10^{-6} cm/s)	Oral absorption class (%)	Number of compounds
Artusson and Karlsson (1991)	Caco-2	>1 ≤ 0.1	100 < 1	20
Yee *et al* (1997)	Caco-2	< 1 1-10 >10	0-20 20-70 70-100	35
Gres *et al* (1998)	Caco-2	> 2.0	100	20
Pade and Stavchansky (1998)	Caco-2	> 10	> 90	10
Stenburg *et al* (2001)	Caco-2	≤ 0.074 ≥ 3.5	10 90	21
Bergstrom *et al* (2003)	Caco-2	≤ 0.2 ≥ 1.6	≤ 20 ≥ 80	27
Hou *et al* (2007c)	Caco-2	≥ 6 .0	High (>80)	69*
Di *et al* (2011)	MDCK II**	~ 3	Low/medium (<80) High (>80)	19
Varma *et al* (2012)	MDCK II**	≥ 5.0	≥ 80/90	97
Pham-The *et al* (2013b)	Caco-2	~ 0.7 ≥ 16 .0	< 30 ≥ 85	324*

*Collated permeability literature values from different laboratories
** MDCKII strain (MDCK-LE) cell line with isolated low endogenous efflux transporter expression

An example from Table 4.1 is by Artusson *et al* (1991), using a dataset of 20 compounds, who defined that a compound would have complete absorption if it had a permeability > 1 x10^{-6} cm/s (Artursson and Karlsson, 1991). More recent studies have indicated higher permeability thresholds than 1 x10^{-6} to define a high absorption compound (Yee, 1997, Stenberg *et al*., 2001, Pade and Stavchansky, 1998). All of the above appear to be based on a subjective selection apart from the recent investigation by Varma *et al* (2012). This study used Receiver Operating Characteristic (ROC) analysis to objectively define the best permeability threshold for fraction absorbed based on a dataset of 82 compounds with permeability measured in a low transporter expression MDCK II cell line. The threshold defined was > 5 x 10^{-6} cm/s for ≥ 80% or ≥ 90% fraction absorbed (Varma *et al*., 2012).

Additionally to Table 4.1, the Food and Drug Administration (FDA) agency has recommended a set of high and low permeability standards compounds with known fraction absorbed (CDER/FDA, 2000). These standard compounds can be measured alongside new chemical entities (NCEs), which are then considered as highly or poorly permeable, depending on whether the permeability is greater or lower than the FDA standards. This can then be related to fraction absorbed based on these

standards. Potential problems with this are the choice of standard. For example, the high permeability standards propranolol, verapamil and metoprolol have differences in their permeability which could result in potential incorrect prediction depending which standard is used when testing alongside NCEs. The recent study by Pham-The *et al*. (2013) established a rank order relationship between Caco-2 permeability and oral absorption for 324 compounds. This was achieved using Caco-2 values collated from different laboratories. The thresholds defined were based on the permeability of the FDA standard metoprolol. In this case, Caco-2 permeability greater than 16×10^{-6} $cm.s^{-1}$, which is 0.8 times the metoprolol permeability was used to take into account the lower HIA threshold of 85% used (Pham-The *et al*., 2013b).

4.6 Most Oral Absorption Models Fail to Take into Account Permeability and Aqueous Solubility

Permeability and solubility are considered the main fundamental properties that govern the rate and extent of oral absorption (Amidon *et al*., 1995). These two important properties are utilised by the Biopharmaceutics Classification System (BCS), which will be discussed further in the next section. For a drug to be absorbed, it must firstly dissolve in the gastrointestinal fluid in order to then permeate the intestinal membrane. As an increasing number of NCEs have high lipophilicity and low solubility, predicting absorption of NCEs is problematic. Inadequate aqueous solubility can lead to poor, erratic, variable absorption, so it is important to consider the effects of solubility for the prediction of intestinal absorption (Miller *et al*., 2011).

In studies by Zhao and co-workers, data with solubility and dose dependency were defined but not used in the majority of the initial oral absorption models. Upon inclusion of compounds with solubility issues, the resulting models had higher error (Zhao *et al*., 2001). It was also noted, however, that the more insoluble a compound, the lower the resulting absorption. In a later study, compounds with no identified solubility issues were used to build models, and some of these resulting models were then used to predict absorption for the compounds with dose-limiting and dose dependency effects. Overall, the prediction of absorption of these excluded

compounds was in agreement with observed values or the models tended to overestimate absorption (Zhao *et al*., 2002).

There is a lack of studies that incorporate solubility into oral absorption models. However, in work by Pham-The *et al*. (2013), oral absorption was predicted, taking into account solubility. In this study, Pham-The, using a rank order relationship, noted that the relationship between permeability and oral absorption is less certain for poorly-absorbed compounds. They also found, using various contour plots, that incorporating solubility improves classification of HIA based on permeability data by about 10%; therefore showing that using solubility in models is potentially advantageous for oral absorption prediction. However, predicted solubility used on its own is not a good predictor of oral absorption, particularly for poorly-absorbed compounds (Hou *et al*., 2007a).

The lack of solubility incorporated into oral absorption models is not surprising given that both permeability and absorption are closely but inversely related with solubility (Buckley *et al*., 2012, Lipinski, 2000). When comparing with permeability, solubility seems not to be regarded as important as permeability in relation to oral absorption; instead, it is regarded as a factor that can lead to poor (solubility limited) absorption, in addition to other limiting factors such as transporter and enzyme effects. Furthermore, the relative importance of solubility could be dependent on the goals of the research organization, and the mechanistic importance of solubility in regards to oral absorption may not be considered (Lipinski, 2000). In spite of this, the main reasons for poor oral absorption have been shown to be either poor permeability or poor solubility or both (Savjani *et al*., 2012).

Even if models were to include experimental solubility, the main issue here is the lack of experimental solubility for drug compounds to be used in oral absorption modelling. Therefore, molecular descriptors that describe the process of solubilisation of the drug, such as crystal lattice energy, solvent cavity formation energy and solvation energy are utilised in the prediction of solubility (Wang and Hou, 2011, Ghafourian and Bozorgi, 2010). The general solubility equation (GSE) is a simple method that predicts aqueous solubility using only two parameters, logP and melting point (Jain and Yalkowsky, 2001). Other methods may employ more

specific molecular descriptors to improve the predictive accuracy (Cheng and Merz, 2003, Chen *et al*., 2002). GSE and its variants have been used for the estimation of oral absorption-related parameters termed absorption potential (Dressman *et al*., 1985, Sanghvi *et al*., 2001). Recently, a melting point based absorption potential (MPbAP) has been proposed, which is derived from the GSE and includes maximum dose, to give an indication of oral absorption. In general, it was found that the lower the melting point the higher the tendency the compound had to be highly-absorbed and *vice versa*. It was also found that for some higher melting point compounds absorption was limited by dose (Chu and Yalkowsky, 2009). Due to the complexity of solubility, it is difficult to find one molecular descriptor to adequately describe all the solubility processes.

In summary, the importance of solubility on oral absorption is highlighted in the literature, but there are few studies that incorporate both experimental solubility and permeability values within a model, in order to see the effect these two properties have on oral absorption (Pade and Stavchansky, 1998, Bergstrom *et al*., 2003).

4.7 There is a Need for Permeability and Solubility Multi-Label Models

The importance of permeability and solubility has been emphasised by their use in the Biopharmaceutics Classification System (BCS) (Amidon *et al*., 1995). The BCS system was developed to classify drugs into one of four classes based on solubility and/or dissolution properties and intestinal permeability (Figure 4.1). The BCS has been adopted by many regulatory authorities as a standard for the justification of biowaivers for costly bioequivalence studies. Compounds that are eligible for biowaivers under the BCS are immediate release dosage forms with high permeability and high solubility (BCS class 1) and are experimentally shown to exhibit rapid dissolution. In addition, the EMA (EMA, 2010) has extended the eligibility of biowaivers to include certain class 3 compounds. Therefore the BCS is shown to be a vital cost effective tool during drug development (Amidon *et al*., 1995, CDER/FDA, 2000).

	High permeability	**Low permeability**
High solubility	Class I	Class III
Low solubility	Class II	Class IV

Figure 4.1. The Biopharmaceutics Classification System (BCS)

In drug discovery the characterization of preliminary BCS classification is of great interest. The use of a provisional BCS class prediction can help guide decision making and formulation of compound development strategies (Ku, 2008, Varma *et al*., 2012, Pham-The *et al*., 2013a, Bergstrom *et al*., 2003, Lennernas and Abrahamsson, 2005, Butler and Dressman, 2010). In addition, it has been observed that knowledge of the different BCS classes can give an indication of the rate limiting steps of absorption as well as potential metabolic routes and transporter interactions (Wu and Benet, 2005, Lennernas and Abrahamsson, 2005).

There are many classification models in the literature that predict oral absorption, solubility or permeability classes in separate models (Ghafourian *et al*., 2012, Gozalbes *et al*., 2011, Gozalbes and Pineda-Lucena, 2010). These classification models predict just one property and assign a compound to one class label out of two or more mutually exclusive class labels, for example high or low absorption. This is single label classification. The problem with this is that in a real life scenario most objects belong to more than one class at the same time. For example a drug molecule can be highly-absorbed but can also have high solubility or low solubility. The prediction of multiple class labels at the same time is termed multi-label classification (Carvalho and Freitas, 2009, Tsoumakas and Katakis, 2007, Read *et al*., 2011). Due to the relationship between solubility and permeability with oral absorption, a potential multi-label problem exists.

Early research into multi-label modelling has focussed on text categorization (McCallum, 1999, Schapire and Singer, 2000) and now this type of method has expanded into being utilised in many different fields such as gene function prediction (Schietgat *et al*., 2010), medical diagnosis (Shao *et al*., 2013), and drug discovery

(Michielan *et al*., 2009). There are two main types of multi-label methods; problem transformation and algorithm adaptation methods. Problem transformation methods involve transformation of the multi-label data into single label data to then carry out conventional single label classification. Therefore problem transformation methods can also be termed algorithm independent methods and be used with any single label classification method. Algorithm adaption methods involve the adaptation of original single-label algorithms to deal with multi-label data directly (Carvalho and Freitas, 2009, Tsoumakas and Katakis, 2007, Read *et al*., 2011).

Problem transformation is a more common route for dealing with multi-label data. There are several different strategies in order to transform multi-label data into single label data for analysis. A common approach is the binary relevance method. This is where each class label, or property, is separately predicted. The results are then combined to give the results for the multi-label problem. In relation to the BCS prediction, solubility and permeability are predicted separately then the predicted BCS is assigned based on the combined permeability and solubility predictions based on the two separate labels. This method is simple and any single label classification algorithm can be used. A benefit of this method is that the compounds in the datasets do not need to be identical as the properties are modelled separately; therefore all available data are used. However, one important drawback of this method is that it fails to take into account label interactions (Carvalho and Freitas, 2009, Tsoumakas and Katakis, 2007, Read *et al*., 2011).

An example of binary relevance multi-label method utilised in the literature for BCS classification is by Pham-The and co-workers (Pham-The *et al*., 2013a). Although the multi-label method termed binary relevance is not mentioned in this study it built separate models for the *in silico* prediction of solubility or Caco-2 cell permeability. The results from the models were then combined to give a provisional BCS prediction (Pham-The *et al*., 2013a). A similar study predicts solubility and rate of metabolism separately to predict biopharmaceutical drug disposition classification class (BDDCS) (Wu and Benet, 2005) using the combined predictions (Broccatelli *et al*., 2012).

Another typical multi-label method in the problem transformation category is called label power set. This is where the two labels to be predicted are converted into a single label by combining the labels (Carvalho and Freitas, 2009). In the context of BCS, this method is basically the prediction of BCS classes directly. Therefore rather than a prediction of solubility and permeability a BCS class is predicted. The only relevant examples in the literature predict BDDCS class (Wu and Benet, 2005, Khandelwal *et al*., 2007), instead of predicting BCS class. In one example the prediction of BDDCS class was carried out using recursive partitioning (building a single decision tree), random forest (building a set of decision trees) and support vector machine (Khandelwal *et al*., 2007). Although this method takes into account interactions between labels, the main problem with this method is the lack of representation of some of the classes. In other words some classes may have fewer examples compared to the rest and leads to a poor predictive accuracy for that underrepresented class (Broccatelli *et al*., 2012). In addition, models can only be built when both labels are known for each compound in the dataset, therefore not utilising all of the data available. Therefore, for this work this method was not utilised due to the drastic reduction of data available for modelling. Note that it is also possible to predict continuous values of permeability and solubility, or another approach would be to classify compounds into multiple categories (low, medium, high)(Macheras and Karalis, 2014). However these approaches are out of the scope of this current work since it concerns in binary classification of chemicals according to the BCS system.

A less well known multi-label method is classifier chain (Read *et al*., 2011). This method seeks to overcome the drawbacks of binary classifier by taking into account label interactions. The method works by firstly predicting one label. Then, the predicted label is used, along with any other predictors (molecular descriptors), in order to predict the second label. Finally, the predictions from both labels are combined like binary relevance for the final BCS prediction. A potential issue with this method could be the noisy data created from using the predicted value of the first label as a descriptor to predict the second label. One of the problems of this method is deciding which label to predict first (Gonçalves *et al*., 2013). In some cases there may be a definite order of the labels from a mechanistic point of view,

making this choice obvious. For example, in the case of solubility and permeability prediction, solubility would be the first label and permeability would be the second. This is because solubility is a basic property that can affect permeability of molecules, whereas permeability is a higher level property. Molecules need to be dissolved and solubilised first, before they can permeate the intestinal wall.

Both binary relevance and classifier chain also require an extra step to convert the single labels into a final label result (BCS class assignment). Both have the benefit of utilising all available data for modelling without being restricted like the power set method. A summary of multi-label methods is presented in Table 4.2

Table 4.2. A comparison of multi-label classification methods.

Method	**Advantages**	**Disadvantages**
Binary relevance (BR)	• Any single label classification algorithm can be used • Simple	• Higher computational cost than power set • Ignores potential label interactions
Label Power set (PS)	• Any single label classification algorithm can be used • Takes into account label interdependences	• Often, there are underrepresented (multi-label) classes with few compounds, which tends to cause over fitting
Classifier chain (CC)	• Takes into account label interdependences	• Which label to choose first? Order of chain has an effect on accuracy (Gonçalves *et al*., 2013) • Noisy data created from using predicted value of the first label

4.8 Summary of the Literature on Oral Absorption Models

From the literature, there are certain problem areas that have been identified as relevant and that could be investigated;

Firstly, oral absorption datasets typically contain many more highly-absorbed compounds than poorly-absorbed compounds, creating highly unbalanced datasets. Models are not reflective of an industry setting and not fit for purpose. Furthermore, there are few methods in the literature that directly cope with datasets with unbalanced class distribution for oral absorption.

Secondly, the main oral absorption models available in the literature are only suitable for the prediction of passively absorbed compounds with no solubility issues due to data exclusions. With the increasing research and the shift in drug candidates towards poorly-absorbed compounds, and the inclusion of compounds undergoing different and multiple transport absorption routes, global models that include all types of absorption mechanisms are sought after.

Thirdly, it is apparent that feature selection is important, and this is highlighted by the large number of oral absorption models utilising some kind of feature selection. Therefore, the impact of feature selection for oral absorption datasets in tandem with methods that cope well with datasets with unbalanced class distribution needs further investigation.

Next, as permeability and solubility are rate limiting steps fundamental to oral absorption, investigation into the effect of these factors into models is lacking. Furthermore the relationship of *in vitro* permeability with absorption needs to be established objectively rather than subjectively.

Finally, the importance of oral absorption has resulted in separate single label models predicting permeability and solubility in the drug discovery literature. However there are few models that predict both permeability and solubility in a multi-label fashion. The area of provisional BCS classification using multi-label methods needs further investigation.

There are a variety of oral absorption models in the literature, which attempt to be interpretable, predictable or both. The review of the literature and all of the above points are important to consider in relation to the aims and objectives of this thesis.

5 Datasets and Methods

This chapter details the generic methodology utilised in this research. Specific information regarding individual methods can be found in individual chapters of this thesis.

5.1 Datasets

A summary of the data available for the four datasets used in this thesis are presented in Table 5.1. This table shows the chapter(s) in which each dataset are used and the number of compounds in each dataset. These datasets including references can be found on the accompanying disk in this thesis.

Table 5.1. Overview of the four experimental datasets used in this thesis

Chapter	Dataset	HIA (%)	Caco-2 Papp (x 10-6 cm/s)	MDCK Papp (x 10^{-6} cm/s)	Aqueous solubility (mg/mL)	Melting Point (^{o}C)	Maximum strength dose
6,7	1	645					
8	2	689					
9	3	931	386	246	482	609	893
10	4		1428	247	750		

5.1.1 Dataset 1

This dataset consisted of Human Intestinal Absorption (%HIA) data for (initially) 647 drugs and drug-like compounds freely available on the internet (http://cadd.suda.edu.cn/admet/) (Hou *et al*., 2007c). These 647 drugs and drug-like compounds covered a wide variety of pharmacological and chemical classes. After the removal of duplicates (sulfamethazine and glycine) a final dataset consisting of 645 compounds was obtained. From the 645 drugs, 95 compounds were identified by Hou *et al* (2007c) to be excluded from the QSAR model development. More precisely, 43 were absorbed via carrier mediated transporters or via the paracellular route, 24 had poor solubility problems, 26 contained ammonium groups, and for 2, logD could not be calculated resulting in a total of 95 compounds. For QSAR modelling it is important that datasets are curated to produce homogenous group of

data with a similar biological mechanism therefore the effect of removing some of these compounds from the model building process was investigated.

5.1.2 Dataset 2

This dataset consisted of the compounds from dataset 1 plus a small collection of additional %HIA data for 47 compounds collected mainly from the dataset of Varma *et al* (Varma *et al*., 2010) and the literature.

5.1.3 Dataset 3

Dataset 3 contained collected data for %HIA, *in vitro* apparent permeability, aqueous solubility, maximum dose strength and melting point.

Human Intestinal Absorption

Using these datasets with collected %HIA as a starting point, an extensive literature search was then carried out and additional compounds were also added from the drug information obtained from the FDA (www.fda.gov) Drugs@FDA database (accessed from June 2012 to May 2013) and the literature. I used the same principles to calculate and evaluate the reliability of fraction absorbed values as defined by other works (Varma *et al*., 2010, Zhao *et al*., 2002). Where there was no numerical value defined in the literature, categorical values for fraction absorbed were also included for this dataset. At the end, this final dataset consisted of 914 numerical and 17 categorical fraction absorption values creating a final dataset of 931 compounds with % HIA data.

In vitro Apparent Permeability

Apparent permeability (P_{app}) data measured in $cm.s^{-1}$ was collected for compounds with known fraction absorption from this dataset. The dataset contains apparent permeability data for two different cell lines, Caco-2 and MDCK, obtained from the literature. The dataset contains 386 Caco-2 and 246 MDCK P_{app} values for drug and drug-like compounds. For 185 compounds the permeability was found for both cell lines, and this dataset was used to investigate the relationship between the two cell lines. Where there were multiple permeability values for a single compound these results were averaged, unless they were very different; in which case comparison of

MDCK and Caco-2 permeability was carried out (if available) or careful examination of the experimental conditions of the specific value was performed in order to justify inclusion.

For Caco-2 permeability, the published dataset by Pham-The *et al* (Pham-The *et al*., 2011) was used as the starting point from which an exhaustive literature search was carried out. For MDCK permeability, permeability data from two studies by Varma and co-workers (Varma *et al*., 2005, Varma *et al*., 2012) were used as a starting point. As there are different strains of this cell line, it was important to reference what strain (if known) was used in the study. In addition, it was decided not to just isolate data collection on one strain, but make a note which would aid in interpretation at a later stage. The main two types of MDCK strains collected were MDCK II and MDCK-MDR1. A preliminary statistical paired t test of these two main strains showed no significant difference between these two strains in this dataset ($p > 0.05$), therefore all the data for MDCK were used together for comparison with Caco-2 (See chapter 9).

Aqueous Solubility

For the compounds with %HIA data, aqueous solubility was collated where available in the literature. Aqueous solubility for 483 compounds in mg/mL was obtained primarily from the AQUASOL dATAbASE (6^{th} Edition) and SRC (PHYSPROP) databases (http://esc.srcinc.com/fatepointer/search.asp) and the literature. For the AQUASOL data, those values that had the highest evaluation codes as defined by the database were selected, and those compounds with more than one value were averaged.

Melting Point

Experimental melting point (in ^{o}C) was obtained from the AQUASOL dATAbASE, SRC (Physprop), the Hazardous substances data bank (HSDB) (http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?HSDB) and the literature. The average was taken if a melting point range was stated.

Maximum Strength Dose

The maximum strength dose was obtained for the compounds in this dataset from the British National Formulary (BNF, 2012), FDA electronic orange book 2012 (accessed December 2012-January 2013) and Martindale (Martindale, 2009). Where there were still missing values, an extensive literature search was carried out and the values presented are the best recommendation based on an evaluation of the literature data. Where doses were based on bodyweight, a body weight of 70kg was used to calculate the maximum dose for human.

5.1.4 Dataset 4

In vitro Apparent Permeability

The permeability dataset used to build the initial permeability models was taken from the published dataset of Pham-The *et al* (2013a). This dataset contained apparent permeability values for 1301 compounds from the Caco-2 cell line, measured in the pH range 6.5-7.4. Upon the removal of duplicates, compounds with incorrect structures and compounds with molecular weights greater than 3000, a dataset of 1288 compounds remained for permeability modelling.

In addition, a selection of *in vitro* permeability data collected from Caco-2 and MDCK cell lines from dataset 3 was used. These 127 compounds were not present in Pham-The *et al*.'s published permeability dataset and had solubility values present. The references for this validation set can be found in Appendix 4.

Aqueous Solubility

Experimental solubility data were obtained from dataset 3 and from the literature. In addition to this, qualitative aqueous solubility was collected for those compounds with missing experimental solubility values. The main source of qualitative solubility was obtained from Martindale (2009) and from the literature. For the 250 qualitative solubility values that were obtained, these were converted to numerical values based on the principles of Kasim *et al*. according to Table 5.2 (Kasim *et al*., 2004). The final total dataset size was 750 compounds.

Table 5.2. Solubility definitions adapted from Kasim *et al*. (2004)

Descriptive term (solubility definition)	**Solubility assigned (mg/mL)**
Very soluble (VS)	1000
Freely soluble (FS)	100
Soluble (S)	33
Sparingly soluble (SPS)	10
Slightly soluble (SS)	1
Very slightly soluble (VSS)	0.1
Practically insoluble (PI)	0.01

5.2 Molecular Descriptors

A variety of different software packages were used to compute 2D/3D molecular descriptors; they include TSAR 3D v3.3 (Accelrys Inc), MDL QSAR (Accelrys Inc.), Advanced Chemistry Development ACD Labs/ LogD Suite v12. For chapter 6, Kowwin (U.S. EPA) was also used. In addition, for chapters 7, 8 and 10, additional molecular descriptors were calculated using MOE (Chemical Computing Group Inc.) v2011.10 and v2012.10. Due to software limitations, some molecular descriptors could not be calculated for some compounds in the datasets.

5.3 Training and Validation Sets

For the majority of datasets the compounds were placed into either training or validation sets randomly after the dependent variable (i.e. %HIA, permeability, solubility) was sorted by ascending values and then by log P. The ascending values were then put into groups depending on the dataset size to ensure that a large and diverse set of compounds were placed into the validation set. For example, for ascending %HIA values, these compounds were put into groups of six, then $5/6^{th}$ of these compounds was placed in the training set and the remaining into the validation set. The details and numbers of compounds in the training and validation sets can be found in the specific chapters of this thesis.

5.4 Feature Selection Techniques

Feature selection can reduce redundancy and chance correlations with molecular descriptors when models are built. Feature selection methods select the best molecular descriptors for the classification or correlation with the dependent variable. In this work feature selection methods are only applied to the training sets or method optimisation sets and never carried out using the compounds in the validation set.

The feature selection methods used in this thesis are summarised in Table 5.3, where an 'x' indicates the feature selection method utilised in the work in that chapter:

Table 5.3. Feature selection methods utilised in this work

	Chapter				Software used
Feature selection method	**6**	**7**	**8**	**10**	
Stepwise regression	x	x			MINITAB v 15.1.0.0
Stepwise Discriminant	x				TSAR v 3.3
Lipinski's rule of five*	x	x			n/a
Predictor importance using random forest			x	x	STATISTICA v 11 and 12
Chi-square			x		STATISTICA v 11
Information gain ratio			x		WEKA v 3.6
Greedy stepwise			x		WEKA v 3.6
Genetic search			x		WEKA v 3.6

*Including number of rotatable bonds

5.4.1 Stepwise Regression Analysis

Stepwise regression analysis was performed on the training sets using MINITAB Statistical Software (version 15.1.0.0) to select descriptors that had significant linear relationships with the %HIA, the dependent variable. The calculated molecular descriptors were set as independent variables. In order to minimise the risk of chance

correlations, the maximum number of descriptors allowed in the models was restricted to eight. All the descriptors selected using this method had a p-value < 0.05, showing that the chosen descriptors by this method were significant for predicting %HIA.

5.4.2 Stepwise Discriminant Analysis

Forward stepwise discriminant analysis was carried out using TSAR 3D v 3.3. This method works by calculating Mahalanobis distance (Mahalanobis, 1936). The independent variable is chosen if it has the greatest increase in the total Mahalanobis distance between the two classes compared to the rest of the molecular descriptors. The seven molecular descriptors selected using this method were only used for classification modelling of HIA class in chapter 6.

5.4.3 Lipinski's Rule of Five Descriptors Plus the Number of Rotatable Bonds.

Lipinski's 'rule of five' is a popular rapid screen to identify compounds that are poorly-absorbed (Lipinski *et al*., 1997). The descriptors proposed by Lipinski are: molecular weight, number of hydrogen bonding donor and acceptor groups and logP. Number of rotatable bonds was also added, as it has been suggested to help predict oral bioavailability and hence oral absorption (Veber *et al*., 2002). This resulted in a total of five molecular descriptors.

5.4.4 Predictor Importance Ranking Using Random Forest (RF)

The molecular descriptor set is generated using random forest. Using STATISTICA v 11 and v 12, random forest generates a set of decision trees based on random subsets of compounds and descriptors in the training set. The ensemble of decision trees vote based on the individual tree results and then the majority vote for a particular compound determines the classification of that compound (Breiman, 2001).

Additionally, random forest calculates an output of predictor importance of all the molecular descriptors used to generate the random forests. A set number of molecular descriptors with the highest ranking predictor importance were selected to be used for model building. The calculation of descriptor importance in

STATISTICA software is explained as follows. For every molecular descriptor, the drop in each node impurity (delta) is summed for all nodes in the trees and expressed relative to the largest sum – i.e. the most significant descriptor. The delta is calculated for every descriptor (even if not used in the node for the splitting of the tree) and summed for every node and tree produced in the forest. The larger the delta the more significant the molecular descriptor is. The final summed delta value for every descriptor is normalized against the most important molecular descriptor and therefore expressed relative to the molecular descriptors with the largest delta. This means that important molecular descriptors that may not have been picked to be in the trees may still appear in the final predictor importance table. Additionally this method allows the application of different misclassification costs for different classes; therefore it can be used to overcome unbalanced class distributions (a problem that occurs in my datasets in general).

Optimization of the random forest method was carried out based on the plot of the misclassification rate vs. the number of trees. The misclassification rate is the number of misclassified compounds divided by the total number of compounds. The lower the misclassification rate, the better the model. Based on the misclassification rate, the optimum number of trees was selected and used to repeat the analysis with the new optimized value. From this optimised model, a set of the top molecular descriptors were selected with the highest predictor importance for each of the dependent variables: HIA (used in chapter 8), permeability and solubility class (used in chapter 10).

5.4.5 Chi Square (CS)

CS is a statistical measure of the association (or dependence) between two categorical variables (Liu and Setiono, 1995). The greater the CS value, the more statistically significant the molecular descriptor is in relation to the %HIA class, therefore allowing the most statistically important molecular descriptors to be ranked. The main drawback of using CS as well as many other filter techniques is that it is a univariate feature selection method; therefore it does not take into account interactions between the molecular descriptors. CS is an association measure for categorical descriptors, therefore there may be problems when continuous variables

are used that contain a large spread of numerical values, since the conversion of numerical variables into categorical ones (required for the use of the chi square measure) may lose relevant information. This feature selection method was carried out using STATISTICA v 11. The software default number of bins (10) was used for chi square discretization of the molecular descriptors. The top 20 molecular descriptors for HIA class were selected using this method in chapter 8.

5.4.6 Information Gain Ratio (IGR)

Information gain ratio is a normalised function of the information gain feature selection method developed by Quinlan as part of the ID3 (Iterative Dichotomiser) decision tree algorithm (Quinlan, 1979). This feature selection method is used to split the decision tree into nodes and identify molecular descriptors that are the best for the individual splits. Information gain works to minimise the information needed to classify compounds into resulting nodes. It is the difference between the original information (before the data are split) and the new information produced after using the molecular descriptor to split the training set data. This difference is the gain of information achieved by using a specific molecular descriptor, therefore the molecular descriptor with the highest gain is the one used for the split. Information gain ratio was first described by Quinlan in the context of the C4.5 algorithm, which superseded ID3 (Quinlan, 1993). The higher the ratio value the better the molecular descriptor for the split. Information gain ratio overcomes the bias towards selecting those molecular descriptors with many numerical values by normalising the information gain. This feature selection technique was carried out using WEKA 3.6 to select the top 20 molecular descriptors for HIA class in chapter 8.

5.4.7 Greedy Stepwise (GRD)

Feature selection methods such as chi square and information gain ratio are based on ranking the molecular descriptors based on a certain criterion and do not take into account the interactions between the molecular descriptors. Greedy stepwise takes molecular descriptor interactions into account as well as the correlation with HIA class. This method seeks to maximise the correlation between HIA and the molecular descriptors being tested, and minimise correlations between the molecular descriptors. Greedy stepwise is a forward stepwise feature selection method,

therefore is a local search method (Kittler, 1978). This local search method firstly considers all the molecular descriptors and picks the best one – i.e., the one that correlates with HIA class. It then starts again with all the remaining molecular descriptors, and picks the best molecular descriptor that best pairs with the previously selected molecular descriptor in relation to HIA class. The iterations carry on until a local maximum is reached. As only a local search can be carried out based on the molecular descriptor(s) selected in all the previous iterations, the potential for a global search of all the different possible subsets is limited, and promising regions of molecular descriptor space can be missed (Dudek *et al*., 2006). The evaluator function used in WEKA v 3.6 to guide the greedy search in the feature selection process was correlation-based feature selection subset evaluator (CfssubsetEval). This evaluator aims to maximise the correlation between the best molecular descriptors and HIA class and also minimises the correlation or redundancy between the descriptors for the search subsets generated. In chapter 8 this method was used for HIA modelling and 21 molecular descriptors were selected and utilised further for model building to predict HIA class.

5.4.8 Genetic Search (GEN)

In chapter 8 this feature selection method is utilised as a filter (rather than wrapper) version of the genetic algorithm (Shah and Kusiak, 2004). Genetic algorithm (GA) was first created by Holland (Holland, 1975), although the concept of genetic algorithm was being researched before this. GA generally is an evolutionary algorithm, which mimics the process of natural evolution. An initial population is created containing random candidate solutions. In the context of this work, a candidate solution is a molecular descriptor subset. Each candidate solution is evaluated in terms of its fitness (quality), and candidate solutions are then selected to be reproduced and to undergo modifications with a probability proportional to their fitness values. The process of selecting "parent" candidate solutions based on fitness and producing "offspring" solutions that are based on the parents is iteratively performed for a number of iterations, so that the population of candidate solutions gradually evolves towards better and better candidate solutions (Holland, 1975). In this work I have utilised the genetic search feature selection method using WEKA software (Goldberg, 1989). This method carries out a global search in the 'molecular

descriptor space' to find the best subset of molecular descriptors relating to HIA class, guided by a subset evaluator that generates a numerical value of 'fitness' (quality) of any given feature subset. Like with the greedy search technique, the evaluation function used for the genetic search method was 'CfssubsetEval'. Using this feature selection method for HIA class, 64 molecular descriptors were selected and utilised further for model building in chapter 8.

5.5 Modelling Techniques

5.5.1 Multiple Linear Regression (MLR)

Linear regression involves deriving a linear model between X and Y variables. Multiple linear regression models the relationship between multiple independent Y variables and the dependent variable, X, by fitting a linear equation to observed data in the training set. MINITAB v15.1.0.0 software was used to carry out MLR in chapter 6. In the context of this work the dependent variable is numerical %HIA and the independent variables are the molecular descriptors as selected via the feature selection methods described in previous sections. The linear equation derived from the compounds in the training set can be used to predict %HIA for the validation set to test the predictability of the MLR model obtained.

5.5.2 Linear Discriminant Analysis (LDA)

LDA is a statistical classification method that finds a linear combination of independent variables that can best characterise or separate two or more classes based on the dependent variable. LDA is different from MLR as the dependent variable is a categorical value rather than a numerical value; however still continuous independent variables are used to build models to predict the categorical class. In chapter 6, LDA, using MINITAB v15.1.0.0, was used to find molecular descriptors that can discriminate between high and low absorption class values using the training set. The models built were then utilized to predict the absorption of drugs in the validation set.

5.5.3 Classification and Regression Trees (C&RT)

STATISTICA v11 and v 12 (StatSoft Ltd) software was used for classification of compounds into classes using C&RT analysis. C&RT analysis is a statistical technique that uses decision trees to solve regression and classification problems developed by Breiman (Breiman *et al*., 1984). If the dependent variable is categorical then a classification tree is made (e.g. predicting low or high absorption classes) and if the dependent variable is continuous then a regression tree is produced, resulting in the prediction of numeric values of the dependent variable.

The binary C&RT analysis starts building the decision tree at the 'tree root' using molecular descriptors. The algorithm in C&RT will choose the most appropriate (statistically significant) molecular descriptor to split the tree and the most appropriate threshold value to define the split. A parent node splits into two child nodes and then these become the parent nodes for the next split. The splitting of the tree continues until it can be no longer split or until the tree satisfies one of the stopping factors being applied by the user to prune the tree to prevent over-fitting. The nodes which cannot be split anymore are termed terminal nodes, and they contain the predicted classes (Breiman *et al*., 1984, Tan *et al*., 2006). In chapters 7-10 in this thesis categorical prediction of binary HIA, permeability and solubility classes was focussed on using C&RT analysis.

5.5.3.1 Classification Thresholds

For classification, when the class variable takes numerical values, a threshold needs to be defined in order to assign compounds into the binary classes. In this investigation, compounds were placed into categorical classes of 'high' or 'low' absorption according to the observed %HIA value in the dataset. For chapters 6-8, the threshold for the classes was 50%; therefore, any compound with %HIA $\geq$ 50% was assigned to the 'High' class, and any compound with a %HIA $<$ 50% was assigned to the 'Low' class. For chapter 9, a range of %HIA thresholds were tested from 30% to 90%. Finally, for chapter 10 specific thresholds for permeability and solubility were assigned, as detailed further in chapter 10.

5.5.4 Multi-Label Classification

Multi-label classification was used to predict multiple class labels simultaneously. In this thesis multi-label classification was used to predict permeability and solubility labels in the form of BCS class prediction using classification C&RT analysis.

For the multi-label work in this thesis, only problem transformation methods were considered, as they are more common for dealing with multi-label data. There are several different strategies in order to transform multi-label data into single label data for analysis (Carvalho and Freitas, 2009); however, in this work the two multi-label methods of binary relevance and classifier chain were used to predict BCS class using permeability and solubility, using dataset 4.

Binary relevance involves building separate models of solubility and permeability, and the predicted BCS is assigned based on the combined permeability and solubility prediction based on the two separate labels (**Figure 5.1**).

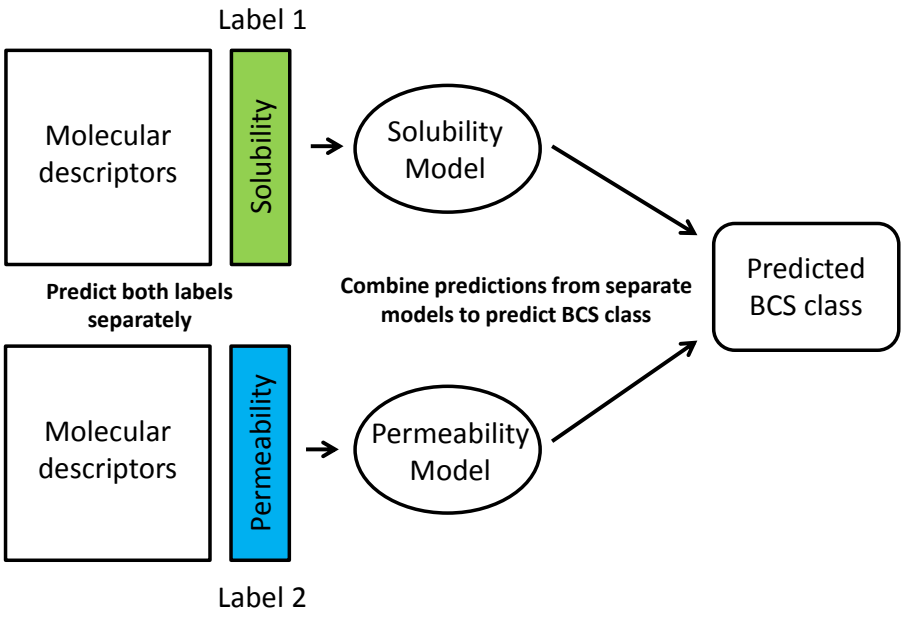

Figure 5.1. How the binary relevance problem transformation method works for BCS class prediction

The second multi-label method utilised in chapter 10 was classifier chain (Read *et al*., 2011). See Figure 5.2 for a schematic representation of this method.

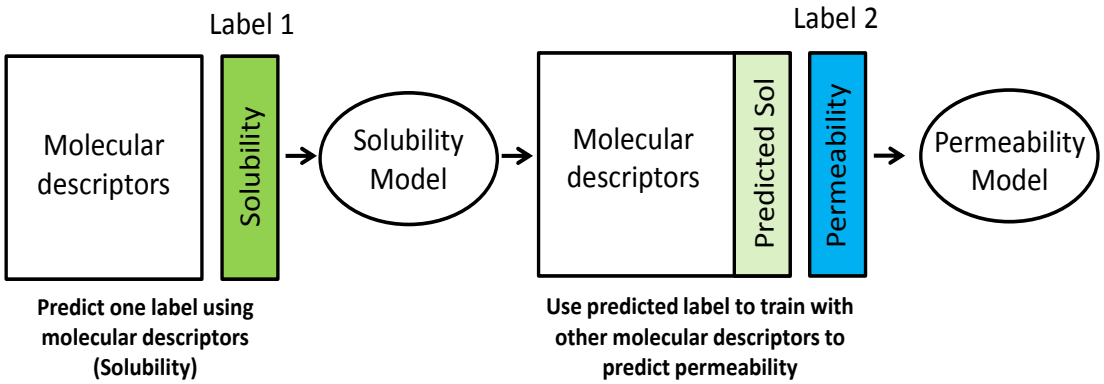

Figure 5.2. Prediction of BCS using the classifier chain multi-label method

The method works by firstly predicting one label (in this case solubility), then using the first predicted label, along with the other molecular descriptors used to build the models, in order to predict the second label (Figure 5.2). Then the predictions from both labels are combined (like in the binary relevance method) for the final BCS prediction.

5.6 Statistical Evaluation of Models

5.6.1 Evaluation of Continuous/Numerical models

For MLR analysis in chapter 6, the following statistical criteria were obtained: N, the number of observations; r^2, the squared correlation coefficient; S, the standard deviation; q^2, leave one out cross validation squared correlation coefficient F, Fisher's criterion. From the predicted and observed %HIA data, the RMSE (root mean squared error) was calculated for the training and validation sets separately (Equation 5.1).

4 / 5 ' L ¾ 5 ' L §Ã:ã å Ø?âÕ;æá **Eq. 5.1**

In equation 5.1, pred is the predicted and obs is the observed %HIA, and n is the number of compounds.

5.6.2 Evaluation of Categorical/Classification models

5.6.2.1 Single-Label Evaluation

For binary classification, the possible outcomes of a prediction model can be visualised using a confusion matrix (Figure 5.3). True positives and true negatives are correct predictions of the classes by the classification model. False positives and false negatives are misclassifications by the model. In the context of oral absorption, a false positive is an observed poorly-absorbed compound that is predicted to be highly-absorbed by the model, and a false negative is a compound that is actually highly-absorbed but is predicted to be poorly-absorbed by the model. The number of correct classifications and misclassifications of both the classes are used in calculations to give an indication of the predictive ability of a classification model.

		Observed class: HIGH	Observed class: LOW
Predicted class	HIGH	True Positive (TP)	False Positive (FP)
	LOW	False Negative (FN)	True Negative (TN)

Figure 5.3. A confusion matrix that outlines the possible outcomes of a binary classification

The predictive performance of classification models was measured using Specificity (SP), Sensitivity (SE), the cost normalised misclassification index (CNMI) and SP × SE.

Specificity is the ratio of correct classifications of poorly-absorbed compounds as depicted by Equation 5.2.

$$SP = TN/(TN+FP) \qquad \textbf{Eq. 5.2}$$

In this equation SP is the total number of true negatives divided by the total of true negatives and false positives as defined in Figure 5.3. Specificity is inversely proportional to the number of false positives.

Sensitivity indicates the correct number of classifications for the highly-absorbed compound class as shown in Equation 5.3.

$$SE = TP/(TP+FN) \quad \textbf{Eq. 5.3}$$

In equation 5.3, SE is the number of true positives divided by the total number of true positives and false negatives predicted by the model. Sensitivity is inversely proportional to the number of false negatives.

In this work, in order to represent the overall predictive performance, specificity multiplied by sensitivity was used (SP × SE)(Parpinelli et al., 2002). Overall accuracy is often defined as the number of correct predictions (true positives and true negatives) divided by the total number of compounds in either the training or validation set (TP + TN)/ (TP + TN + FP + FN)). However, this calculation is not suitable to use for this work when the dataset has a highly unbalanced class distribution, especially for %HIA data. In other words, due to the majority of highly-absorbed compounds in the training and validation sets, the classification outcome of these compounds disproportionately affects the overall accuracy: therefore, accuracy will follow the same trend as the sensitivity values in the model and fail to take into account the specificity appropriately. For example, if a dataset contained 90% of highly-absorbed and 10% of poorly-absorbed compounds, a trivial classifier would consist of predicting the highly-absorbed class (the majority class) for all compounds in the validation set. Such a trivial majority classifier, which does not involve any data analysis, would trivially achieve an overall accuracy of 90% (if accuracy is simply measured as (TP + TN)/ (TP + TN + FP + FN)). However, this high accuracy is misleading. Although the majority classifier achieved perfect prediction for the high absorption class, it achieved no correct predictions for the poorly-absorbed class. This example clearly shows a weakness of the overall accuracy measure, which is not an appropriate measure to use when the class distribution is very unbalanced. The use of SP × SE avoids the above problem in this scenario, since the trivial majority classifier would achieve a prediction of 0% by multiplying the sensitivity (100%) and specificity (0%), and this would highlight the majority classifier's poor ability to classify both classes. A measure of 0% accuracy for the majority classifier is also intuitively fair; since that classifier is not even

taking a look at the value of the descriptors (it just counts the number of compounds in each class in order to determine the majority class). In summary, the overall accuracy as defined in the literature is an incorrect measure of accuracy in problems with very unbalanced class distributions, like the oral absorption datasets used in this thesis. In this work, to overcome the problem, the better accuracy measure of SP × SE was used as the measure of the overall accuracy. This measure is a better representation of a model's predictive power, as it is affected the same way by false negatives as by false positives, and therefore was used in measuring the overall accuracy of classification models in this thesis.

The cost normalised misclassification index (CNMI) was calculated using equation 5.4. This is a useful evaluation measure when different misclassification costs are applied to false positives and false negatives in models predicting HIA, permeability and solubility classes in chapters 7, 8 and 10.

The CNMI is calculated in the following way: The numerator of this equation is calculated by first multiplying the number of each type of misclassifications (false positives and false negatives) by the corresponding misclassification cost and then adding those two products. The denominator (normalization factor) is calculated by first multiplying the total number of compounds in each class – i.e. number of negatives (poorly-absorbed compounds) and number of positives (highly-absorbed compounds) – by the corresponding misclassification costs and then adding those two products.

$$\%0/+L = \frac{:\hat{o}\,\text{þH}\,\tilde{n}\quad \hat{o}\,\text{þ};>:\hat{o}\,\text{üH}\,\tilde{n}\quad \hat{o}\,\text{ü};}{:\text{ü}\quad \text{H}\,\tilde{n}\quad \hat{o}\,\text{þ};>:\text{þ}\quad \text{H}\,\tilde{n}\quad \hat{o}\,\text{ü};} \qquad \textbf{Eq. 5.4}$$

$Cost_{FP}$ and $Cost_{FN}$ are the misclassification costs assigned for false positives or false negatives; Neg is the total number of poorly-absorbed compounds and Pos is the total number of highly-absorbed compounds. Note that the numerator of equation 5.4 is the total misclassification cost obtained by using a classification model to classify compounds in the training or validation set, whilst the denominator is the maximum misclassification cost that could in principle be achieved (if all compounds in the validation set were misclassified). Hence, the calculated value will be between zero and one, with zero representing no misclassifications, as the number increases to one

then the misclassifications of the model increase. The value of this measure has been calculated for models predicting HIA, permeability and solubility classes in chapters 7, 8 and 10.

5.6.2.1 Multi-Label Evaluation

For chapter 10, the evaluation of multi-label classification models requires different measures from conventional single label classification models (Tsoumakas and Katakis, 2007, Carvalho and Freitas, 2009). The statistical evaluation of multi-label work can be difficult, as a result can be fully correct, partially correct or fully incorrect. Therefore, it is important to have several different evaluation measures, due to the issue of multiple class labels, to help select the best model, i.e. the one with the best model performance over a set of evaluation measures.

For multi-label classification there are two broad types of evaluation measures. These are label based evaluation measures and label set evaluation measures (Tsoumakas and Katakis, 2007, Read *et al*., 2011, Carvalho and Freitas, 2009). Label based evaluation measures are those based on the individual single labels, such as Hamming loss (Schapire and Singer, 2000) and classification/subset accuracy (McCallum, 1999, Zhu *et al*., 2005). In this work, the accuracy of the individual four BCS classes was used, which is essentially the converse of the Hamming loss – in the sense that the latter is to be minimized, whilst the individual accuracy per class is to be maximized. The individual class accuracy for each class was calculated by dividing the correct number of predictions for compounds of that class divided by the total number of compounds of that class, resulting in four accuracy measures for the individual four BCS classes. Additionally, for this work the SP X SE accuracy measure of the individual permeability and solubility labels was calculated.

Label set evaluation measures are based on the prediction of all labels together. Therefore this type of measure can be very harsh, as if there is not a perfect prediction of both labels for a compound, that prediction will be considered completely wrong, even if one of the two labels was correctly predicted. Examples of label set evaluation measures are micro-averaging and macro-averaging (Yang, 1999). The label set evaluation measures used in this work are based on macro-

averaging (Yang, 1999). Macro-averaging is the average, by compound, of all the accuracies for the different BCS classes. To calculate the overall accuracy, the number of correct predictions (regardless of class) was divided by the total number of compounds. It must be noted that the overall accuracy calculated in this way could be biased and not give an accuracy measure which would show the predictive accuracy of all four classes. Therefore, in addition the geometric mean of all four individual predictive accuracy measures for the BCS classes was calculated. The geometric mean is calculated by multiplying all the four BCS class accuracy measures and taking the fourth root of this product. The benefit of this measure is that it will not be biased towards the distribution or predictive accuracy of any individual BCS class. In other words, if a model can predict three out of four classes with high accuracy but is unable to predict accurately for one class, the geometric mean accuracy will be low.

6 The Effect of Under-Sampling the Majority Class of the Training Set for Oral Absorption Models

6.1 Introduction

There are many oral absorption models that are based on unbalanced class-distribution datasets (Zhao *et al*., 2001, Wessel *et al*., 1998). In oral absorption datasets, there are typically many more highly-absorbed than poorly-absorbed compounds. This creates biased models, which are better at predicting the majority highly-absorbed compounds compared with the minority poorly-absorbed class. This defeats the objective of a good model, which should be able to predict both high and low absorption with high level of accuracy. Furthermore, models produced will not reflect real life drug discovery scenarios where a higher proportion of drugs are more likely to be poorly-absorbed (Wu and Benet, 2005).

There are resolutions to this problem of biased models due to unbalanced class datasets. One potential resolution is the method of under-sampling. Under-sampling involves removal of compounds of the highly-absorbed majority class from the training set to create a balanced training set of 50:50 distribution of high and low absorption compounds (Figure 6.1).

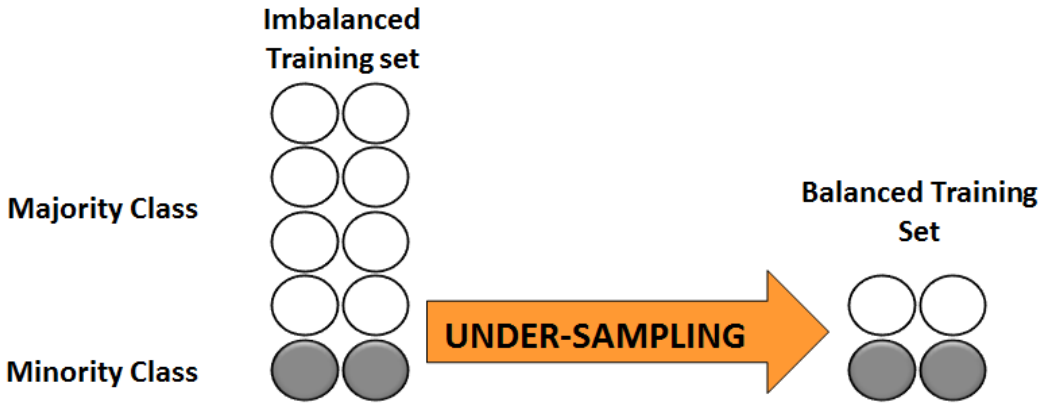

Figure 6.1. Under-sampling the majority class in an unbalanced class-distribution dataset to create a balanced training set for modelling

Although studies have attempted to resolve the problem of unbalanced class-distribution datasets, there is little evidence of work comparing the predictive ability

of models built using unbalanced vs and balanced data using the under-sampling method for oral absorption (For examples of recent studies see (Talevi *et al*., 2011, Bai *et al*., 2004)).

Therefore, this chapter addresses the problem of unbalanced class distributions by creating a balanced training set through under-sampling the highly-absorbed compounds. In doing so, this chapter seeks to achieve more effective models with a better predictive accuracy for the poorly-absorbed compounds without jeopardising the predictive accuracy for the highly-absorbed drugs.

6.2 Methods

6.2.1 Dataset

Dataset 1 was utilised for the prediction of oral absorption as defined in the 'Dataset and Methods' section 5.1.1 of this thesis. In this study, the 26 compounds containing ammonium groups were excluded entirely to avoid the added complications as stated in the methods section 5.1.1. These 26 compounds were a part of the 95 compounds as identified by Hou *et al* (Hou *et al*., 2007c). Therefore in this chapter the statement "Upon exclusion of the 95 compounds" means the exclusion of the remaining 69 compounds as defined by Hou *et al* (Hou *et al*., 2007c), as the 26 compounds containing quaternary compounds have already been removed. The remaining 619 compounds where %HIA was available were used in the resulting analyses.

6.2.2 Training Sets and Validation Sets

The dataset was split into a training set and a validation set. From the same original dataset, different training sets were created and had different numbers of compounds. From the remaining compounds, either the rest were used as a validation set or a selection of the remaining compounds was used as a validation set. Table 6.1 is a summary of the numbers of compounds assigned to the training and validation sets for the different splits of the datasets. The next section describes the rationale behind partitioning the dataset into these training and validation sets.

Table 6.1. Datasets of intestinal absorption and numbers of compounds in each set (n)

Training Set Name	Validation Set Name	n Training Set	n Validation set	n Total
TS1	VS1	94	502	596
TS2	VS2	496	100	596
TS1 or TS2	VS3	94 or 496	89	183 or 585

6.2.2.1 TS1-VS1

The overall dataset contained more highly-absorbed than poorly-absorbed compounds, especially in the 80-100 %HIA range. Therefore, random selection of compounds into training and validation sets could result in a higher number of the highly-absorbed compounds, creating a bias towards this majority. To overcome this issue, a balanced dataset was devised. Compounds were sorted by ascending %HIA, then by ascending log P values. From the sorted dataset, 10 categories of %HIA, each of which contained about 10 drugs from each 10% range of %HIA, were created; and for each category, compounds were selected randomly for the training set. By taking 10 samples from each %HIA range, under-sampling the majority class is achieved due to the larger number of highly-absorbed compounds being removed and not used compared with the poorly-absorbed class. The remaining compounds were used as the validation set (VS1). Under-sampling the majority class (high absorption compounds) gave a more balanced training set with similar numbers of low and high absorption drugs in the training set.

6.2.2.2. TS2-VS2

The dataset was initially sorted based on ascending %HIA and then by ascending logP values. Then, from each group of six consecutive compounds, five were assigned to the training set, and one compound was allocated to the validation set randomly. This ensured similar distributions of %HIA values in the training and validation sets. The resulting dataset is unbalanced and not under-sampled, and it has a %HIA distribution similar to the one in the original dataset, with a higher proportion of highly-absorbed compounds in the training and validation sets.

6.2.2.3 VS3

An additional balanced validation set containing 89 compounds was used to compare the models developed from the training sets, TS1 and TS2. This is because it would be unfair to compare models from these different training sets as the validation sets are not identical. Recall for the TS1 training set, the initial validation set, VS1 (of 502 compounds), consisted of all the remaining compounds not used in the training set. Therefore the training and validation sets have different %HIA distributions. In other words, not only the validation sets of TS1 and TS2 are very different in terms of the number of the compounds, for TS1, the validation set is not a correct representation of the training set as the %HIA distributions are very different.

Therefore a new validation set, VS3, containing 89 compounds was developed as follows. From the remaining 502 compounds in VS1, 89 compounds were selected randomly by under-sampling the highly-absorbed compounds (VS3). None of the compounds from TS2 were included in this validation set. This new validation set had a similar %HIA distribution of 50:50 highly:poorly-absorbed compounds to that of the TS1 training set. This new validation set enabled direct comparisons between all the models comprehensible when using the results from the validation set VS3.

6.2.2.4 Exclusion of Outliers

The removal of the 95 compounds as highlighted by Hou *et al*. (2007c) from the dataset reduced the number of compounds in the training and validation sets. The final numbers left in the training and validation sets after these compounds were removed was for TS1, 73 and 477; and for TS2, 458 and 92 respectively. Removing the outliers did not affect the balance of high to low absorption compounds significantly for VS1, VS2, VS3 or TS2. For TS1 the balance changed towards highly-absorbed compounds from the initial 50:50 split to 33:67.

6.2.3 Model Development

6.2.3.1 Molecular Descriptors

A total of 215 descriptors were used in this study using a variety of different software including TSAR 3D (Accelrys Inc.), MDL QSAR (Symyx Inc.), Kowwin (U.S. EPA) and Advanced Chemistry Development ACD Labs/LogD Suite v 12.

6.2.3.2 Feature Selection

For this chapter the following feature selection methods were used to select molecular descriptors:

1) Stepwise regression analysis
2) Stepwise discriminant analysis (used only for LDA classification analysis)
3) Lipinski's rule of five plus number of rotatable bonds

The molecular descriptors selected by these methods were used in the models developed by regression and discriminant analysis. There were a significant number of compounds that had missing values for descriptors such as ACD Density, therefore stepwise regression was carried out again excluding these descriptors and a second model was developed for the TS1.

6.2.3.3 QSAR Modelling Techniques

The following QSAR methods were used in this chapter:

1) Multiple linear regression (MLR) was used for the continuous prediction of %HIA.
2) Linear discriminant analysis (LDA) for the categorical prediction of HIA class.

For MLR analysis, predictive models were built using observed %HIA set as the dependent variable and each set of the molecular descriptors selected by stepwise regression analysis and the rule of five descriptors as independent variables.

For the LDA analysis, according to observed %HIA values in the dataset, compounds were placed into either the "high" class if %HIA was equal to or greater

than 50% or the "low" class if %HIA was less than 50%. In this manner, predictive classification models were developed using the observed %HIA class as the response, and each set of the descriptors selected by stepwise regression analysis, stepwise discriminant descriptors and the rule of five descriptors. Molecular descriptors selected via stepwise discriminant analysis were solely used for the classification of the compounds into highly-absorbed (%HIA ≥ 50%) or poorly-absorbed groups (%HIA < 50%) and not for prediction of precise %HIA values using regression analysis.

6.3 Results

In this work, continuous and categorical models were developed using different training sets with different data distributions, using subsets of molecular descriptors. The names of the molecular descriptors and a brief description of each descriptor used in the models in this chapter can be found in the Appendix 1 (Table A1.1).

6.3.1 Regression Models

Two regression models were developed for the training set TS1; the under sampled training set contains roughly a 50:50 distribution of high:low absorption compounds (Equations 6.1 and 6.2). These models were obtained using the descriptors selected by stepwise regression when all the descriptors were used in analysis (model 1) and when several descriptors with a high number of missing values (ACD_density and logP) were excluded (model 2). The statistics reported are r^2, squared correlation coefficient; q^2, cross validation coefficient; F, Fisher's criterion; S, standard deviation and n, number of compounds.

Model 1 Stepwise Regression 1 TS1 (Equation 6.1)

%HIA = 125 - 0.357 SHHBd - 0.627 SHBint2 + 4.71 ACDLogD5.5 - 0.00643 Inertia Moment 2 Size - 0.516 SHBint7 - 297 SpcPolarizability - 22.2 ACD_Density - 1.24 SsCH3

n = 94 S = 15.7 r^2 = 0.755 q^2=0.690 F = 32.7 **Eq. 6.1**

Model 2 Stepwise Regression 2 TS1 (Equation 6.2)

%HIA = 101 - 0.0753 ACD_PSA + 4.02 ACDLogD7.4 - 2.72 ka3 - 0.272 SHBint2 - 6.16 aliphatic rings(5) - 2.98 SHBint2_Acnt - 284 SpcPolarizability - 0.275 SHBint3

n = 94 S = 16.1 r^2 = 0.742 q^2=0.687 F = 30.5 **Eq. 6.2**

Using TS2, which is a randomly selected training set of 496 compounds, stepwise regression model 3 was obtained. Model 4 is the regression equation obtained for TS2 using Lipinski's 'rule of five' parameters.

Model 3 Stepwise Regression 3 TS2 (Equation 6.3)

%HIA = 95.4 - 0.138 ACD_PSA - 12.9 ACD_Rule_Of_5 - 3.22 ACDLogD2 - 1.35 SHBint9 + 6.27 ACDLogD5.5 + 3.48 SdsssP

n = 496 S = 16.1 r^2 = 0.686 q^2=0.674 F = 178.2 **Eq. 6.3**

Model 4 – Ro5 Descriptors (Ro5) TS2 (Equation 6.4)

%HIA = 98.5 + 0.0072 Mass - 1.08 Rotatable Bonds - 5.12 H-bond Donors - 2.40 H-bond Acceptors + 2.34 ACD_LogP

n = 496 S = 20.2 r^2 = 0.533 q^2=0.510 F = 112.0 **Eq. 6.4**

From the original dataset by Hou *et al* (2007c) there are 95 compounds that were excluded for a variety of reasons, as previously mentioned in the methods section 5.1.1. These remaining outliers were removed from the dataset and regression analysis was performed again. It must be noted that only some of the outliers fell within the training sets and the remaining belonged to the validation set. Additionally, for a better comparison of the models, RMSE values were calculated for the new common validation set, VS3, containing 89 compounds for all 4 models. Table 6.2 shows the statistical parameters of the equations obtained for the training sets before and after the exclusion of the outliers. Table 6.2 also indicates the

average prediction error (RMSE) for the validation sets and for the new validation set, VS3.

Table 6.2. Statistical parameters and prediction accuracies of regression models for training (t) and validation (v) sets

Model (E.q.)	Training Set Name	Validation Set Name	r^2	F	S	RMSE		n	
						t	v	t	v
6.1	TS1	VS1	0.755	32.73	15.66	14.90	25.49	94	502
6.2	TS1	VS1	0.742	30.51	16.08	15.29	26.11	94	502
6.3	TS2	VS2	0.686	178.2	16.56	17.30	20.37	496	100
6.4	TS2	VS2	0.533	111.1	20.17	20.37	23.24	496	100
Common Validation set (VS3)									
6.1	TS1	VS3	0.755	32.73	15.66	14.90	25.05	94	89
6.2	TS1	VS3	0.742	30.51	16.08	15.29	24.45	94	89
6.3	TS2	VS3	0.686	178.2	16.56	17.30	30.83	496	89
6.4	TS2	VS3	0.533	111.1	20.17	20.37	38.64	496	89
After exclusion of 95 compounds (Hou et al. 2007c)									
6.1	TS1	VS1	0.788	29.80	15.41	14.43	24.40	73	477
6.2	TS1	VS1	0.785	29.25	15.52	14.54	23.84	73	477
6.3	TS2	VS2	0.697	172.9	15.37	15.24	18.59	458	92
6.4	TS2	VS2	0.540	106.0	18.91	18.79	22.38	458	92

r^2-correlation coefficient; F-Fisher's criterion; S-standard deviation; RMSE-root mean squared error; n-number of compounds, t-training set; v-validation set

From Table 6.2, the most suitable equation based on the statistics for the training set was equation 6.1, which used dataset TS1. Equation 6.1 shows a slightly better r^2 to the training set than equation 6.2 using the same training set. In Table 6.2, equations 6.3 and 6.4 (developed using TS2) appeared to have poorer statistics for the training sets; however, their RMSE for the validation sets is better than equations 6.1 and 6.2. However, it must be noted that a direct comparison of equations 6.1 and 6.2 with equations 6.3 and 6.4 at this point is not coherent. This is because models 6.1 and 6.2 were derived from a small training set of 94 compounds (TS1) and evaluated using VS1, a large validation set (n=502), with a different %HIA distribution to the training set; whereas models 6.3 and 6.4 were derived using a much larger training set (TS2) and were evaluated using VS2, a smaller validation set that was representative of the training set in terms of %HIA distribution.

For a better comparison of the models with each other, a new validation set of 89 compounds (VS3) was randomly selected from the original large validation set of TS1. This new validation set was assembled in a way that the %HIA distribution was similar to the TS1 training set, i.e., similar numbers of compounds at the different %HIA ranges. RMSE values were calculated for the predicted %HIA values obtained from equations 6.1-6.4 for the 89 compound in VS3, the new validation set. The results in Table 6.2 show that model 6.2 has the lowest RMSE value of 24.45. Models 6.3 and 6.4 have much larger RMSE values for the new validation set, indicating that these models work well for estimation of %HIA of highly-absorbed compounds as shown by the results in Table 6.2, but the estimation accuracy is dropped when the validation set consists of roughly equal proportion of highly and poorly-absorbed compounds, which may be true in real life drug-candidates. This is expected due to the highly biased nature of the training set used for the development of these models (TS2).

In Table 6.2 there is not much difference between the RMSE and r^2 values of the models before and after the exclusion of the compounds identified by Hou *et al* as being absorbed actively or whose absorption is believed to be dissolution limited. When these remaining compounds were excluded from the models above in Table 6.2, the statistics were improved slightly for both the training and validation sets for all models.

6.3.2 Classification Models

Stepwise discriminant analysis selected seven descriptors for the classification of %HIA class using TS1. The descriptors were: number of six-membered aromatic rings, ACD $LogD_{5.5}$, fraction of drugs ionised at pH 1, SdsssP_acnt, SHBint3, SHBint7 and SHHBd (Appendix 1, Table A1.1). Following this, discriminant analysis was performed using the %HIA class as defined in the main methods section and molecular descriptors selected by the three stepwise regressions on TS1 or TS2 (descriptors in Eqs 6.1-6.3), Lipinski's rule of five descriptors plus the number of rotatable bonds, and descriptors selected by stepwise discriminant analysis (total of five models). Tables 6.4 and 6.5 show the measures of predictive

accuracy (measured on the training and validation sets) of the discriminant models for TS1 and TS2, respectively.

Table 6.3. Results of Discriminant Analysis Models using training set TS1 and measured on the validation set VS1

Model	Set	Accuracy (SP X SE)	Sensitivity (SE)	Specificity (SP)	Descriptor Set
1	TS1	0.736	0.943	0.780	Stepwise Regression (Eq. 6.1)
	VS1	0.719	0.915	0.786	
2	TS1	0.757	0.887	0.854	Stepwise Regression (Eq. 6.2)
	VS1	0.752	0.879	0.855	
3	TS1	0.597	0.906	0.659	Stepwise Regression (Eq. 6.3)
	VS1	0.913	0.959	0.952	
4	TS1	0.671	0.887	0.756	Lipinski Rule of 5 plus number of rotatable bonds
	VS1	0.753	0.904	0.833	
5	TS1	0.814	0.962	0.846	Stepwise Discriminant Analysis
	VS1	0.685	0.871	0.786	
After exclusion of 95 compounds (Hou et al. 2007c)					
1	TS1	0.727	0.918	0.792	Stepwise Regression (Eq. 6.1)
	VS1	0.785	0.968	0.811	
2	TS1	0.748	0.898	0.833	Stepwise Regression (Eq. 6.2)
	VS1	0.675	0.925	0.730	
3	TS1	0.626	0.939	0.667	Stepwise Regression (Eq. 6.3)
	VS1	0.920	0.973	0.946	
4	TS1	0.689	0.918	0.750	Lipinski Rule of 5 plus number of rotatable bonds
	VS1	0.804	0.959	0.838	
5	TS1	0.689	0.918	0.750	Stepwise Discriminant Analysis
	VS1	0.756	0.932	0.811	

t-training; v-validation; Sensitivity is equivalent to the number of correctly classified highly-absorbed compounds and is calculated using SE=(TP/(TP+FN)); Specificity is equivalent to the number of correctly classified poorly-absorbed compounds and is calculated using SP=(TN/(TN+FP)); TP-true positive; FN-False negative; TN-true negative; FP-false positive; Overall Accuracy of the models was calculated by multiplying Specificity by Sensitivity (SP x SE).

Table 6.3, referring to models built from training set TS1, shows that for the classification of the validation set the best overall classification accuracy as measured by SP x SE was 0.913, which equated to 481/502 correct predictions using model 3. This model had the highest specificity value of 0.952 (40/42) and the second best sensitivity of 0.959 (441/460) for VS1 validation set. However, for this model, the overall accuracy and specificity are much lower for the training set compared with the validation set. In fact, in most cases the accuracy, specificity and sensitivity of many models are better for the validation set than for the training set.

This can be due to the compound composition of the training and validation set, with the training set containing, by chance, more outlier compounds. This could also be due to the biased distribution of %HIA of the compounds and lack of poorly/moderately absorbed compounds represented in validation set.

Considering also the training set, the best model taking into account the overall accuracy, specificity and sensitivity values for the training and validation sets was model 1. However, there were many descriptor values missing (ACD_Density and logP), therefore this model may not be appropriate in a real life setting as these descriptors may be difficult to obtain for new compounds. From this perspective, the best applicable model considering the training and validation sets is model 2. This achieved an overall SP X SE of 0.751 resulting in 522/596 total correct predictions, and with specificity and sensitivity values of 0.879 (451/513) and 0.855 (71/83) respectively, when those measures are calculated over all compounds in the full dataset (merging the training and validation sets). Model 3 has better SP x SE accuracy of 0.769 (556/596) when calculated this way; however, models 1 and 2 have better overall accuracies for the training set than model 3, as mentioned previously.

Table 6.4. Results of Discriminant Analysis Models using training set TS2 and measured on the validation set VS2

Model	Set	SP X SE	SE	SP	Descriptor Set
1	TS2	0.712	0.958	0.743	Stepwise Regression (Eq. 6.1)
	VS2	0.538	0.942	0.571	
2	TS2	0.722	0.937	0.771	Stepwise Regression (Eq. 6.2)
	VS2	0.591	0.919	0.643	
3	TS2	0.718	0.967	0.743	Stepwise Regression (Eq. 6.3)
	VS2	0.598	0.930	0.643	
4	TS2	0.739	0.958	0.771	Lipinski Rule of 5 plus number of rotatable bonds
	VS2	0.471	0.942	0.500	
5	TS2	0.737	0.956	0.771	Stepwise Discriminant Analysis
	VS2	0.414	0.965	0.429	
After exclusion of 95 compounds (Hou et al. 2007c)					
1	TS2	0.739	0.985	0.750	Stepwise Regression (Eq. 6.1)
	VS2	0.538	0.988	0.545	
2	TS2	0.739	0.985	0.750	Stepwise Regression (Eq. 6.2)
	VS2	0.700	0.963	0.727	
3	TS2	0.695	0.982	0.708	Stepwise Regression (Eq. 6.3)
	VS2	0.605	0.951	0.636	
4	TS2	0.655	0.982	0.667	Lipinski Rule of 5 plus number of rotatable bonds
	VS2	0.450	0.988	0.455	
5	TS2	0.692	0.977	0.708	Stepwise Discriminant Analysis
	VS2	0.538	0.988	0.545	

t-training; v-validation; Sensitivity is equivalent to the number of correctly classified highly-absorbed compounds and is calculated using SE=(TP/(TP+FN)); Specificity is equivalent to the number of correctly classified poorly-absorbed compounds and is calculated using SP=(TN/(TN+FP)); TP-true positive; FN-False negative; TN-true negative; FP-false positive; Overall Accuracy of the models was calculated by multiplying Specificity by Sensitivity (SP x SE).

The classification results obtained for TS2 (Table 6.4) indicate that the classification of poorly-absorbed drugs (specificity values) are less accurate than the highly-absorbed compounds (sensitivity). Moreover, the specificity values of the models developed using TS2 are much lower than models developed using TS1 (compare Tables 6.3 and 6.4). This is due to the unbalanced training set used (TS2), with a lower number of poorly-absorbed compounds compared to highly-absorbed compounds. On the other hand, sensitivity is higher in most models obtained for TS2, compared with models developed with TS1, with exceptions being the validation set sensitivity of model 3 and model 5.

For both TS1 and TS2 the effect of removal of the excluded compounds as highlighted by Hou *et al* (2007c) increased overall accuracy, specificity and sensitivity values in the majority of cases, but there was not a significant increase of evaluation metrics when comparing with and without excluded compounds. So, in practice leaving these compounds in will achieve a more applicable model that will have better generalization for new compounds.

In order to compare the models, the accuracy (SP x SE), sensitivity and specificity of discriminant analysis for the classification of the new validation set, VS3, (containing 89 compounds) using all 5 models were calculated and the results are in Tables 6.5 and 6.6 for TS1 and TS2, respectively.

Table 6.5. Discriminant Analysis Results for new validation set (VS3) for TS1

Model	Set	SP X SE	SE	SP	Descriptor Set
1	VS3	0.677	0.880	0.769	Stepwise Regression (Eq. 6.1)
2	VS3	0.643	0.760	0.846	Stepwise Regression (Eq. 6.2)
3	VS3	0.816	0.860	0.949	Stepwise Regression (Eq. 6.3)
4	VS3	0.657	0.800	0.821	Lipinski Rule of 5 plus number of rotatable bonds
5	VS3	0.551	0.717	0.769	Stepwise Discriminant Analysis

t-training; v-validation; Sensitivity is equivalent to the number of correctly classified highly-absorbed compounds and is calculated using SE=(TP/(TP+FN)); Specificity is equivalent to the number of correctly classified poorly-absorbed compounds and is calculated using SP=(TN/(TN+FP)); TP-true positive; FN-False negative; TN-true negative; FP-false positive; Overall Accuracy of the models was calculated by multiplying Specificity by Sensitivity (SP x SE).

Table 6.6. Discriminant Analysis Results for new validation set (VS3) for TS2

Model	Set	SP X SE	SE	SP	Descriptor Set
1	VS3	0.689	0.918	0.750	Stepwise Regression (Eq. 6.1)
2	VS3	0.724	0.878	0.825	Stepwise Regression (Eq. 6.2)
3	VS3	0.696	0.898	0.775	Stepwise Regression (Eq. 6.3)
4	VS3	0.718	0.898	0.800	Lipinski Rule of 5 plus number of rotatable bonds
5	VS3	0.681	0.939	0.725	Stepwise Discriminant Analysis

t-training; v-validation; Sensitivity is equivalent to the number of correctly classified highly-absorbed compounds and is calculated using SE=(TP/(TP+FN)); Specificity is equivalent to the number of correctly classified poorly-absorbed compounds and is calculated using SP=(TN/(TN+FP)); TP-true positive; FN-False negative; TN-true negative; FP-false positive; Overall Accuracy of the models was calculated by multiplying Specificity by Sensitivity (SP x SE).

Comparing the sensitivity and specificity values reported in Tables 6.5 and 6.6 for the consistent validation set of 89 compounds (VS3), it can be seen that the best specificity is achieved for models developed using TS1 (Table 6.5) with the highest

value obtained using model 3. On the other hand the best sensitivity values were obtained using models developed using TS2 (Table 6.6). This shows that TS1, the balanced dataset has a better classification accuracy compared with TS2 (unbalanced dataset) for predicting poorly-absorbed compounds; whereas TS2 has a better classification accuracy for highly-absorbed compounds due to the biased nature of the dataset.

6.4 Discussion

The aim of this study was to create models that could predict %HIA values or classify compounds into highly and poorly-absorbed classes with emphasis on using a training set with a balanced class distribution to see if there was an improvement in prediction and classification accuracy for poorly-absorbed compounds. Data splitting techniques such as the Kennard-Stone algorithm (1969), which is used for the training set selection based on the molecular descriptor values, could have been used to split the data; however, it would not be useful in this case since the highly-absorbed compounds cover a much larger descriptor space compared with poorly-absorbed compounds (Tropsha 2010). For instance, it was observed in this dataset that splitting the data based on the logP values, or based on the first or the second principal components produced by principal component analysis using all molecular descriptors, will still select for the training set a majority of highly-absorbed compounds (85%-86% highly-absorbed). Therefore, to overcome the well-documented problem of highly biased training sets (towards the highly-absorbed compounds), a subset of the available data was selected with under-sampling of the majority group (highly-absorbed compounds). The selected training set consisted of about ten compounds in each 10% %HIA ranges (94 in total). The models generated using this training set (TS1) were compared with the models generated for a randomly selected training set consisting of 5/6 of the dataset entries (TS2). Regression analysis and classification analysis results are discussed highlighting the best model for each type of analysis.

6.4.1 Regression Models

Regression models were generated using the biased training set, TS2 and the under-sampled training set, TS1. The best overall model with the lowest RMSE for the new validation set VS3 is equation 6.2.

Equation 6.2 contained the following descriptors: PSA, log $D_{7.4}$, Ka3, SHBint2, aliphatic rings(5), SHBint2_Acnt, SpcPolarizability and SHBint3. All of these descriptors can be used in combination to correlate with intestinal absorption; however the correlation decreases significantly when the descriptors are used independently, highlighting that absorption is a complex process and is reliant and influenced by a number of different descriptors, not just one (Hou *et al*., 2007a).

PSA has been found to be the most popular descriptor used in the prediction of intestinal absorption, since its first use in relation to brain penetration (van de Waterbeemd and Kansy 1992). It is a measure of the area of the Van der Waals surface that arises from oxygen and nitrogen atoms or hydrogen atoms bound to these atoms. PSA is related to hydrogen bonding capacity, which is one of the main influencers of passive drug absorption along with lipophilicity (Palm *et al*., 1997). PSA is used more frequently and is deemed more suitable than normal hydrogen bonding potential descriptors as it accounts for the 3D effects of the molecule, such as shielding of the polar functional groups by other atoms. It has been shown that if a molecule has a dynamic PSA of ≥ 140Å, it is likely to have poor absorption (<10% HIA) and if the molecule has a PSA ≤60Å, %HIA values >90% can be achieved (Palm *et al*., 1997, Clark, 1999). This is in agreement with my results as PSA has a negative impact on absorption. Hydrogen bonding ability has been further characterised in model 6.2 by three topological descriptors of SHBint2_Acnt, SHBint2 and SHBint3, all of which have negative coefficients in agreement with the literature.

Hydrophobicity is another physiologically important parameter for intestinal absorption. Molecular descriptors relating to hydrophobicity, such as logP and logD, have a positive contribution to the predictions of passively absorbed compounds (Hou *et al*., 2007c) An increase in hydrophobicity would initially increase the permeation of the compound into and through the intestinal membrane. However, it

has been suggested that the relationship between logD and permeation is non-linear, so if the drug is too hydrophobic with a very high logD value, for example, it may not penetrate the membrane at all and can affect other physicochemical properties such as solubility (Comer, 2003, Varma *et al*., 2010, Kerns and Di, 2008). On the other hand, if a compound has a low logD value this could also prevent absorption unless the compound is small enough (MW less than 200 Da) to be absorbed via the paracellular route (Stenberg *et al*., 2000, Martinez and Amidon, 2002). In this work, $logD_{7.4}$ has been used in model 6.2, which is the apparent octanol/water partition coefficient at pH 7.4. This particular descriptor has been used in other studies, as well as other logD values at lower pH values (Hou *et al*., 2007b, Hou *et al*., 2007c). It has been indicated that although logP is easier to calculate from structure, logD has a better prediction ability, as it takes into account the pH and ionisation. Studies have also shown that a combination of PSA and logD have good prediction abilities for intestinal absorption, indicating that it is a combination of descriptors that influence predictions (Hou *et al*., 2009, Hou *et al*., 2007a)

There are numerous summary tables in the literature compiling an overview of results for previously published oral absorption models (Hou *et al*., 2006, Suenderhauf *et al*., 2011, Talevi *et al*., 2011). This enables a comparison between the results obtained here and those from previous studies in the literature. However, it must be emphasized in this chapter and throughout this thesis that it is very difficult to compare these models, due to the lack of compound information regarding data distribution for the training and validation sets and lack of consistency in validation techniques (Stouch *et al*., 2003). The only way this possibly could be done would be to mimic the datasets used and compare the models on previous works' datasets (Davis and Brunea, 2003).

The results presented in Table 6.2 shows that there is not much difference between the RMSE and r^2 values of the models before and after the exclusion of the compounds that are believed to be absorbed actively or whose absorption are dissolution limited. This indicated that the effect of transporters does not have a significant effect on the goodness of fit of the regression models and has been shown in other work in the literature (Talevi *et al*., 2011). The reason for the small

difference could be because for some compounds, although known to be absorbed via transporters, this process may not be dominant and the effect of the transporters is insignificant compared with passive diffusion of the compounds (Sugano *et al*., 2010, Smith *et al*., 2014). So, in practice, leaving these compounds in may be more realistic and help build generic models with a variety of absorption mechanisms, rather than removing these compounds and possibly reducing the applicability of the model (Suenderhauf *et al*., 2011).

Taking my best regression model, which was model 2 using dataset TS1 with exclusions, an r^2 value of 0.785 was achieved with a RMSE for the training and validation set VS1 of 14.54 and 23.84, respectively. Other studies that used regression analysis such as Wessel *et al* (1998), Zhao *et al* (2001) and Niwa *et al* (2003) are comparable to my model with regards to the training set. However, the RMSE for my validation set is slightly higher, apart from Niwa *et al* (2003). Wessel *et al* (1998) achieved small RMSE values of 9.5% and 16% for the 76 and 10 compounds used in the training and validation sets. Zhao *et al* (2001) with an r^2 value of 0.83 achieved a RMSE of 14%. However, Zhao *et al* only had 38 and 131 compounds in the training and validation sets, respectively. The more recent study by Niwa *et al* (2003) showed that, although a small RMSE value of 6.5% was achieved for the training set, a much larger RMSE value of 27.7% was obtained for the validation set. The numbers of compounds in the training and validation set were 67 and 9 compounds, respectively.

The studies mentioned so far have used small datasets, and so it might not be suitable to compare some of the models in this chapter with models in those studies. Moreover, the comparison of the validation set compounds and distribution of compounds in them is not known. In fact, Klopman *et al* (2002), who used a larger dataset with 417 and 50 compounds in training and validation sets, achieving a r^2 value of 0.79 which is comparable to my model, highlighted that the dataset was limited and that it covered limited chemical space, even with an increase in the number of compounds in the dataset. Therefore, comparing RMSE values without considering the number of compounds used is not appropriate, as the chemical space of the training set and applicability of the models to a wider variety of chemicals are

different (Klopman *et al*., 2002). Therefore, models reported using small datasets may not be as applicable when databases expand further to include new structurally diverse compounds of the future.

The more recent studies carried out by Hou *et al* (2007 and 2009) and Yan *et al* (2008) both used 647 compounds but then excluded the 95 outliers for their work. Yan *et al* (2008) created 3 partial least squares (PLS) models using 380 and 172 compounds for the training and validation sets. The RMSE value of the best model in this study was 18.18%. The best published method is by Hou *et al* (2007c), which achieved r^2 values of 0.90 and 0.84 and RMSE values of 7.8% and 11.2% for the training and validation sets, respectively, using genetic function approximation (GFA). Finally the study by Talevi *et al* (2011) utilised a balanced training set (n=90 compounds) to build linear and non-linear models. Upon exclusion of outliers, i.e. those undergoing carrier-mediated absorption, they achieved an r^2 of 0.663 for the training set. Although comparable, this r^2 value is much lower than the best results obtained in this chapter.

6.4.2 Classification Analysis

There are many advantages of using regression models for the prediction of intestinal absorption. However, depending on the stage at which the prediction models are used, the need for precise values predicted by a regression method may be questionable, when classification methods can be used to define which drugs will be highly-absorbed and therefore more likely to be orally administered compared with those poorly-absorbed compounds which are more likely to be administered via other routes due to poor absorption (Suenderhauf *et al*., 2011).

For the classification analysis, in order to classify which compounds would be grouped as highly-absorbed or poorly-absorbed, a cutoff of 50% of the %HIA value was defined. The choice of 50% was arbitrary, although it has also been used in previous studies (Niwa, 2003). There have been a number of different HIA cutoffs used to group compounds into either highly or poorly-absorbed groups in the literature; these range from 30% (Hou *et al*., 2007c) up to 70% (Xue *et al*., 2004), with no standard defined. One thing to note is that example HIA thresholds used in

the literature will result in different ratios of highly and poorly-absorbed compounds in dataset. Therefore, there will be separate classification problems, each with a different level of difficultly. Therefore, the threshold chosen can significantly affect the model results.

As stated previously, other studies in the literature have compiled summary tables that detail the accuracy, specificity and sensitivity of previously published classification work (Suenderhauf *et al*., 2011, Hou *et al*., 2006, Talevi *et al*., 2011). Overall a similar pattern emerges that the overall accuracy and sensitivity values of previous studies are higher than the specificity values obtained. This could be due to the low ratio of poorly-absorbed compounds in the training sets. For example, Perez *et al* (2004) used linear discriminant analysis to classify a dataset of 209 compounds with training and validation set of 82 and 127 compounds, respectively. This research created two models, one which focussed on classification of %HIA using a threshold of ≤30% HIA and the other focussing on classification using a threshold of >80% HIA. Both training and validation sets are heavily biased towards highly-absorbed compounds. Higher sensitivity values of 0.955 and 0.835 and much lower specificity values of 0.765 and 0.722 were obtained with the threshold of ≤30% and >80%, respectively (Perez *et al*., 2004). An exception to this pattern is the results obtained by two studies by Hou *et al* (2007a,c), where specificity values in the validation set were higher than the sensitivity values. For the results presented in this chapter, the overall accuracy and sensitivity are comparable or higher than previous studies apart from Hou *et al* (2007c), which used the same dataset but excluded carrier mediated and poorly soluble compounds. Additionally, for one study Hou *et al* included the 26 compounds with positively charged nitrogen which are known to be poorly-absorbed and predicted readily with a count of positively charged nitrogen atoms. This simple rule aided the statistics of their model by increasing the specificity (Hou *et al*., 2007c). Additionally, the use of a 30% HIA threshold to define between highly and poorly-absorbed compounds could be an easier classification problem to model using the complex non-linear methods of recursive partitioning, genetic function approximation (GFA) and Support Vector Machines (SVM) (Hou *et al*., 2007a, Hou *et al*., 2007c). The study carried out by Perez *et al* (2004) also indicates that using the lower threshold of 30% better model

predictability was achieved (Perez *et al*., 2004). Again, as mentioned earlier, it would be risky to take results at face value without considering the real impact of information such as the number of compounds in each class in the training and validation sets.

As stated previously in the literature review, according to the literature there is a lot of business emphasis on reducing the number of false negatives in drug discovery. However, there is also a cost effective incentive to reduce false positives. With this in mind, comparing the sensitivity and specificity values reported in Tables 6.5 and 6.6 for the consistent validation set of 89 compounds (VS3), it can be seen that the best specificity is achieved for models developed using the balanced training set TS1 (Table 6.5) with the highest value obtained using model 3. On the other hand the best sensitivity values were obtained using models developed using TS2 (Table 6.6). This shows that TS1, the balanced dataset, has a better classification accuracy compared with TS2 (unbalanced dataset) for predicting poorly-absorbed compounds, whereas TS2 has a better classification accuracy for highly-absorbed compounds due to the biased nature of the dataset. In relation to the reduction of false positives and false negatives, depending on the priority, the balanced TS1 dataset would aid to reduce false positives by increasing specificity and TS2 would increase sensitivity and therefore reduce false negatives. In conclusion, if reducing the number of false positives is the priority, then under-sampling of the majority class of the highly-absorbed compounds would lead to more accurate and applicable *in silico* models for use in industry.

6.5 Conclusion

In this chapter, the dataset of Hou *et al* (2007c) was used for the development and validation of the models. In order to improve the predictive accuracy for the poorly-absorbed compounds, the training set was selected by under-sampling the highly-absorbed compounds. Two types of linear methods were used for the development of the models: linear discriminant analysis for the classification and multiple linear regression for the regression type analysis.

In terms of the linear regression models, results were conclusive that using the balanced dataset with similar proportions of various %HIA ranges leads to more

robust models with lower prediction error for the validation set. This is despite the lower number of compounds in this training set (N=94), in comparison with the randomly selected training set of 496 compounds. It is interesting to note that the r^2 values of this study are comparable to some of the models obtained using a variety of more complex techniques such as SVM and GEN feature selection, showing that simple regression can obtain just as good r^2 and fit for the prediction of %HIA (Reynolds *et al*., 2009, Yan *et al*., 2008).

The discriminant models for the classification of compounds into high and low absorption classes indicated that the use of the balanced training set significantly improves specificity of the models, indicating the higher accuracy of the classification of poorly-absorbed compounds. However, the sensitivity of the models developed using the balanced training set was lower than the sensitivity of the models based on the randomly selected training set, which is skewed towards the highly-absorbed compounds. Therefore, it can be suggested that, for reducing the number of false positives, it is better to use the balanced training set, despite the smaller training set size due to the under-sampling of the majority class.

To conclude, this work highlights that, by creating a training set with a balanced class distribution, improved models which are also applicable to real life scenarios can be achieved for both regression and classification type analyses. It is envisaged that this conclusion may be extended to models based on more complex statistical techniques such as non-linear methods to improve the predictive accuracy further. Even though different models were developed in this chapter, there were particular descriptors that were in more than one model. These descriptors help and confirm the understanding of the process of oral absorption. Descriptors such as logD, PSA and those involving H bonding are all known to have an impact, whether this is positively or negatively, on oral absorption. Another significant point that needs to be considered in training set selection, in future research, is the impact of solubility and the potential distribution of solubility values in real life datasets. Taking this into account may lead to even more applicable models, given the increasing number of the poorly water-soluble and high molecular weight NCEs.

7 Coping with Unbalanced Class Oral Absorption Datasets Using Under-sampling and Misclassification Costs

7.1 Introduction

As previously stated, oral absorption datasets are biased due to the higher number of highly-absorbed compounds compared with poorly-absorbed compounds in the datasets. In the previous chapter the impact of under-sampling the training set to overcome the unbalanced class distribution to produce more accurate, industry-applicable models was investigated. The problem with under-sampling is the reduction in data utilised for model building, therefore there could be a problem with generalization to new compound sets. Additionally, in order to assess the predictability of the balanced training set fairly, the validation set should also be under-sampled to mirror the training set in terms of distribution of the classes, but again this reduces the validation set size and could potentially increase the variability of the results (Blagus and Lusa, 2010). In spite of this, the models built using this balanced class distribution should be better models to predict both poorly and highly-absorbed compounds if a big enough dataset is used.

Another potential resolution to overcome unbalanced class datasets is to increase the cost of misclassification of the minority class. Recall the possible outcomes of a binary classification, as shown in Figure 7.1a. There are two types of misclassification, false positives and false negatives. A poorly-absorbed compound misclassified into the highly-absorbed class would be a false positive (FP), and a highly-absorbed compound misclassified into a poorly-absorbed class would be a false negative (FN).

The ultimate goal of classification models is to predict correctly which is computationally achieved by the algorithms that try to classify in a way that minimises the cost arising from misclassifications. By applying a higher misclassification cost for false positive or false negatives, a higher cost is then associated with the specified misclassification and the algorithm avoids this misclassification due to its higher costs. This should result in a higher accuracy for

the class whose misclassification was assigned the higher cost and should improve overall accuracy. An example of this is shown graphically in Figure 7.1b.

a)

		Observed class: HIGH	Observed class: LOW
Predicted class	HIGH	TP	FP
	LOW	FN	TN

b)

		Observed class: HIGH	Observed class: LOW
Predicted class	HIGH	NO COST	2
	LOW	1	NO COST

Figure 7.1. a) A binary classification matrix showing predictive outcomes b) Binary classification matrix with higher misclassification cost assigned to false positives

According to Figure 7.1b, if the algorithm attempts to misclassify the poorly-absorbed compound into the highly-absorbed class (a misclassification at the intersection of the 'HIGH' row and 'LOW" column in Figure 7.1b), there will be a higher cost associated with this misclassification in comparison with the misclassification of a highly-absorbed drug into the poor absorption group. By increasing the cost for misclassification in this example to two, the number of false positives should be reduced more than the number of false negatives.

There are numerous oral absorption models in the literature; but the research topic of under-sampling and misclassification costs has not been fully explored. This chapter investigates the use of under-sampling and higher misclassification costs to overcome unbalanced class distributions using decision trees (a type of non-linear classification method). Firstly, the use of under-sampling to create a balanced training set will be compared with the use of an unbalanced training set, when building a model using decision trees. Secondly, the effect of applying higher misclassification costs to the balanced and unbalanced datasets will also be investigated. Furthermore the unbalanced dataset will have higher misclassification costs applied to reduce false positives. Therefore, the aims of this chapter are to see the effect of misclassification costs on a balanced training set model and to determine if misclassification costs can be helpful in overcoming imbalanced class distribution in oral absorption datasets, in order to produce more accurate and more applicable models for use in drug discovery.

7.2 Methods

7.2.1 Dataset

Dataset 1 was utilised for the prediction of oral absorption as defined in the 'Dataset and Methods' section 5.1.1 of this thesis. In this work only the 26 compounds containing a quaternary ammonium were removed entirely due to a number of missing molecular descriptors significant to absorption, such as logD, for these compounds.

7.2.1 Training and Validation Sets

Two training sets and corresponding validation sets were selected from this dataset; training set 1 (TS1) containing roughly a 50:50 ratio, and training set 2 (TS2) containing roughly an 85:15 ratio of highly and poorly-absorbed compounds. The same class distribution for the corresponding validation sets was applied to create a fairer more controlled validation for the models. The exact compound numbers and class distributions are shown in Table 7.1.

Table 7.1. Compound numbers and class distribution for both training set scenarios

	Number of Compounds		**Class Distribution (Ratio of High/Low absorption compounds)**	
Dataset	**Training set**	**Validation set**	**Training set**	**Validation set**
TS1	94	89	50:50	50:50
TS2	517	102	85:15	85:15

7.2.2 Model Development

7.2.2.1 Molecular Descriptors

A total of 215 descriptors were used in this study using a variety of different software including TSAR 3D (Accelrys Inc.), MDL QSAR (Symyx Inc.), Kowwin (U.S. EPA) and Advanced Chemistry Development ACD Labs/LogD Suite v 12.

7.2.2.2 Feature Selection

For this chapter as well as using all 215 molecular descriptors for development of models, the molecular descriptors as selected using stepwise regression analysis using training sets TS1 and TS2 and molecular descriptors derived from Lipinski's rule of five plus number of rotatable bonds (four descriptor sets) from chapter 6 were used.

7.2.2.3 QSAR Modelling Techniques

Classification analysis was carried out using C&RT analysis in STATISTICA v 11 (StatSoft Ltd). According to observed %HIA values in the dataset, compounds were placed into either the 'High' class if %HIA was equal to or greater than 50% or the 'Low' class, if %HIA was less than 50%.

For this work, HIA Class was set as the dependent categorical variable and all 215 molecular descriptors were selected as continuous independent variables. Furthermore, pre-selected subsets of descriptors were used in the analysis from the previous chapter. Molecular descriptors were: 1) those chosen by linear stepwise regression for TS1 (also used in Eq. 6.1), 2) those chosen by linear stepwise regression for TS1 (also used in Eq. 6.2), 3) those chosen by linear stepwise regression for TS2 (also used in Eq. 6.3), and 4) descriptors of Lipinski's rule of five including number of rotatable bonds from the previous chapter (Chapter 6).

During C&RT analysis, models were created using descriptor sets 1 and 2 for TS1 and descriptor set 3 for TS2. The validation set was never used at any stage of model development and remained intact for the validation for all of the models. Moreover, Lipinski's 'rule of five descriptors' was used in C&RT analysis for both TS1 and TS2.

The stopping factors applied for building models using TS1 and TS2 were minimum number of cases (compounds) for splitting the data at 10 and 30, respectively. Stopping factors prevent the C&RT method from further splitting the current decision tree if there are fewer compounds than the stopping factor specified in the node. If there were any trees with only one compound in a terminal node, manual

pruning was carried out to prevent over fitting. All other settings used were default settings defined by the software.

7.2.2.4 Misclassification Costs

By applying higher misclassification costs to certain misclassifications (either false positives or false negatives) it is possible to reduce the number of misclassifications in the class with the higher cost. This chapter compares the use of the same costing with higher costing to reduce either false positives or false negatives.

The cost assigned to the misclassification can be subjective. However, to assign a number objectively, the class distribution of the high and poorly-absorbed compounds of the training set should be considered. For TS1, the balanced dataset, a misclassification cost of two was applied to either reduce false positives or false negatives. As TS2 is unbalanced due to the class distribution of the dataset towards the highly-absorbed compounds (85:15), a misclassification cost ratio of 4:1 was applied to false positive:false negatives. It must be noted that due to the class distributions for TS2 the dataset is already biased towards reducing false negatives, as there are more highly-absorbed compounds than poorly-absorbed ones; therefore higher misclassification costs do not need to be applied to false negatives, but they need to be applied to false positives.

7.3 Results

Predictive models for the classification of drug candidates into high and poor absorption groups are very useful in drug discovery. Unbalanced distribution of data in the available datasets has been a drawback which has traditionally complicated the model development activities. In this chapter, two different training sets with different data distributions and various misclassification costs were used to develop classification trees using the C&RT analysis. In all result tables in this chapter the highest SP, SE, SP × SE and the lowest CNMI for the validation sets are highlighted in **bold**. When comparing the models it must be noted that the most significant molecular descriptors selected for splitting the data by the C&RT algorithm will be affected by the class distribution of the training sets, so for TS1 and TS2 with different class distributions different significant descriptors could be picked.

Moreover, when comparing models developed using the same training set, CNMI maybe a more suitable performance measure since it is normalised for the cost ratios of false positives and false negatives.

7.3.1 C&RT Classification Analysis for TS1

Classification using C&RT analysis was carried with the same or different misclassification costs to reduce either false positives or false negatives for TS1, the under-sampled balanced training set. Initially all 215 molecular descriptors were set as independent variables and the HIA class was set as the dependent categorical variable. In this way the C&RT algorithm selects the most significant descriptor out of all 215 for each split. These trees were compared with C&RT trees created by using smaller descriptor sets selected previously by stepwise linear regression using TS1 (descriptor sets 1 and 2), TS2 (descriptor set 3) or descriptors related to Lipinski's rule of five plus number of rotatable bonds (descriptor set 4), as described in the previous chapter.

Table 7.2 shows the predictive performance measures of the classification trees for TS1 obtained with different misclassification costs using all descriptors and descriptor sets 1-4. Recall that SE, SP and SP × SE measures should be maximized, whilst the CNMI measure should be minimized.

Table 7.2. The results of C&RT classification analysis using different descriptor sets and misclassification costs ratios for TS1

Model	Cost FP:FN	Descriptor Set	Set	n validation set	SP × SE	SE	SP	CNMI
1	1:1	ALL	t	83	0.899	0.981	0.917	0.045
			v		0.598	0.733	0.816	0.229
2		1	t	89	0.939	0.962	0.976	0.032
			v		0.625	0.714	0.875	0.213
3		2	t	89	0.951	1.000	0.951	0.021
			v		0.657	0.796	0.825	0.191
4		4	t	89	0.828	0.943	0.878	0.085
			v		0.300	**0.857**	0.350	0.371
5	2:1	ALL	t	83	0.962	0.962	1.000	0.014
			v		0.404	0.667	0.605	0.352
6		1	t	89	0.939	0.962	0.976	0.027
			v		0.547	0.592	**0.925**	0.188
7		2	t	89	0.981	0.981	1.000	0.007
			v		0.604	0.755	0.800	0.203
8		4	t	89	0.920	0.943	0.976	0.034
			v		0.597	0.796	0.750	0.217
9	1:2	ALL	t	83	0.872	0.981	0.889	0.048
			v		0.635	0.778	0.816	0.223
10		1	t	89	0.885	0.981	0.902	0.044
			v		**0.686**	**0.857**	0.800	**0.165**
11		2	t	89	0.951	1.000	0.951	0.015
			v		0.657	0.796	0.825	0.209
12		4	t	89	0.829	1.000	0.829	0.052
			v		0.438	0.796	0.550	0.295

FP = False positive; FN = False negative; t-training; v-validation; Sensitivity is equivalent to the number of correctly classified highly-absorbed compounds and is calculated using SE=(TP/(TP+FN)); Specificity is equivalent to the number of correctly classified poorly-absorbed compounds and is calculated using SP=(TN/(TN+FP)); TP-true positive; FN-False negative; TN-true negative; FP-false positive; Overall Accuracy of the models was calculated by multiplying Specificity by Sensitivity (SP x SE); n, validation is the number of validation set compounds that was predicted by the model CNMI = Cost normalised misclassification index

In Table 7.2, a cost ratio of 2:1 for FP:FN indicates that a double misclassification cost has been applied for the misclassification of poorly-absorbed compounds compared with the misclassification of highly-absorbed compounds, and so forth. Therefore, in this case, the expectation is a reduction in the number of false positives (increased specificity).

In order to see the effect of cost ratios, one should compare the performance measure values of the models generated using the same descriptor set. It can be seen in Table 7.2 that when all descriptors were used in the analysis (models 1, 5 and 9), better predictive accuracy is obtained when misclassification costs are adjusted to reduce false negatives (model 9). In this case the SP × SE increased from 0.598 in model 1 to 0.635 in model 9 and the sensitivity was the highest at 0.778. The CNMI also decreased from 0.229 (model 1) to 0.223 (model 9). This indicates that by applying costs to reduce false negatives a more accurate C&RT model has resulted. The decrease in false negatives (higher sensitivity value) was expected as misclassification costs were adjusted to improve the class prediction of highly-absorbed compounds; however the specificity decreased. The classification tree (model 9) has been presented in Figure 7.2.

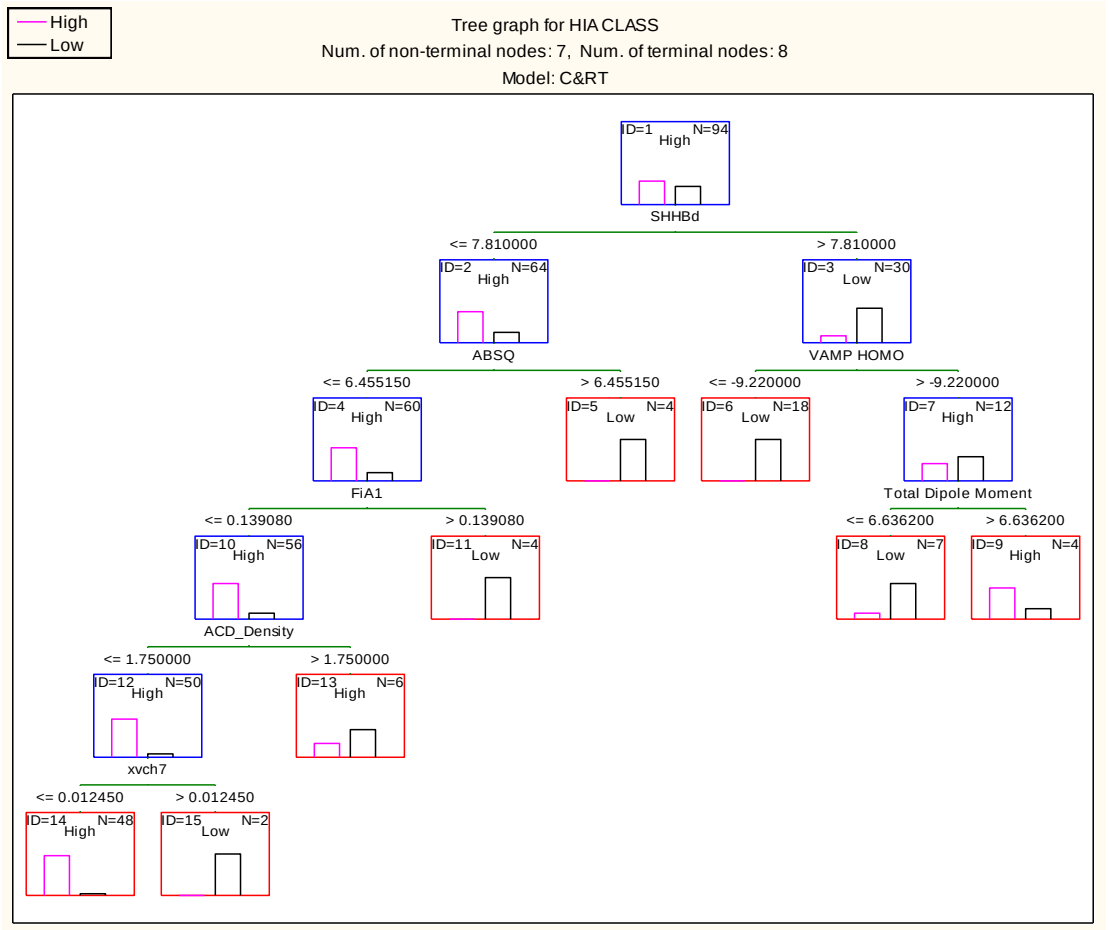

Figure 7.2. Tree graph for the best C&RT model selecting all molecular descriptors using TS1 training set with misclassification costs applied to reduce false negatives (Model 9)

Furthermore, using Table 7.2 one can compare the above results with the results for equal and higher misclassification costs for models built using the molecular descriptors chosen by linear stepwise regression for the estimation of %HIA (descriptors sets 1 and 2) and descriptors of Lipinski's rule of five including number of rotatable bonds (descriptor set 4). Table 7.2 shows that the model obtained using descriptor set 1 is the best model (model 10). The fact that most models that are obtained using a pre-selected descriptor set have better predictive accuracy indicates that such descriptor selection methods may be better than the embedded descriptor selection algorithm in C&RT. Model 10 achieved an SP × SE of 0.686, sensitivity value of 0.857 and a specificity value of 0.800 when using a cost ratio of 1:2 for FP:FN. This model is shown in Figure 7.3. It is interesting to note in Table 7.2 that specificity is much better with the model obtained with higher cost for the false positives, using descriptor set 1, which shows that the misclassification costs are having the expected effect on the model.

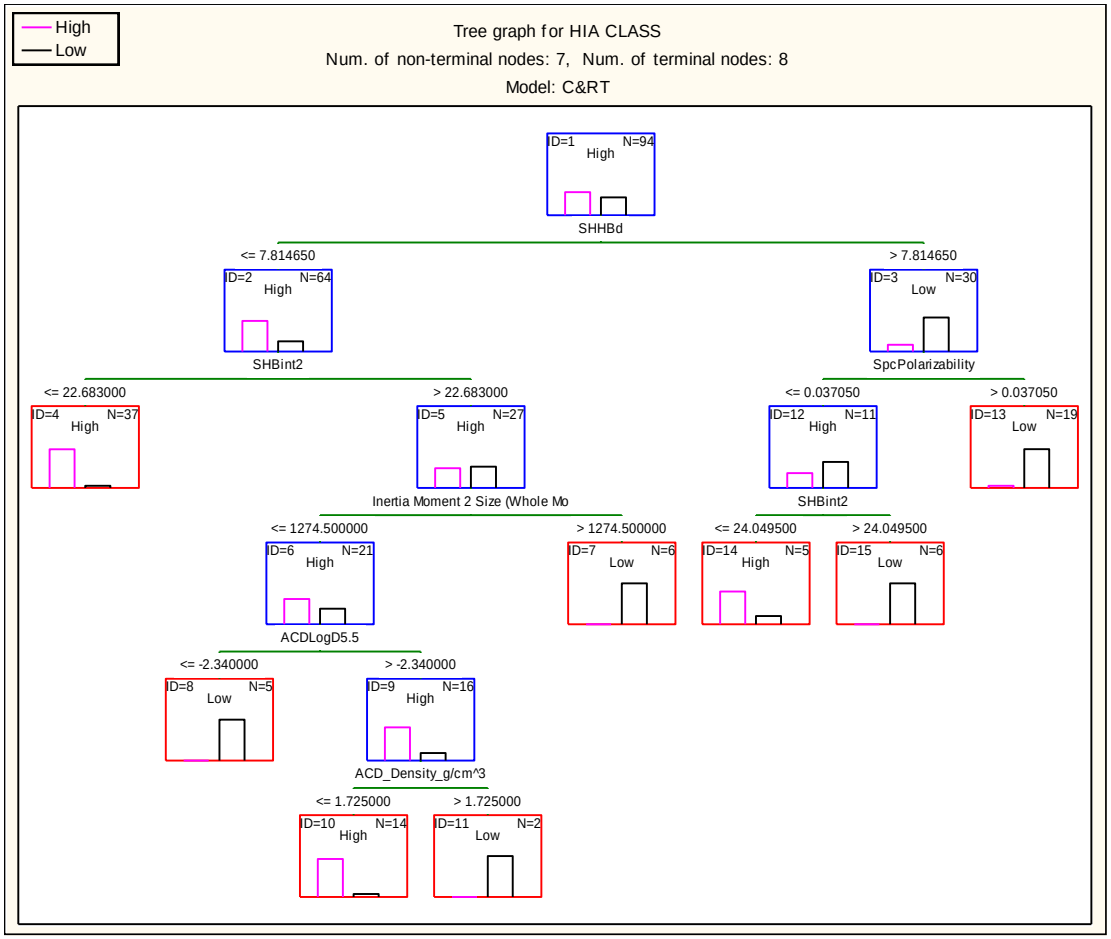

Figure 7.3. Tree graph for C&RT analysis using TS1 with misclassification costs applied to reduce false negatives using descriptor set 1 (Model 10)

There is a general pattern when misclassification costs are applied to either reduce false positives or false negatives for the majority of models (Table 7.2). When higher misclassification costs are applied to reduce false positives, the specificity values are higher or equal to models where similar costs are applied, with only a few exceptions. On the other hand, false negative values decrease upon assigning higher misclassification costs on false negatives, resulting in higher or equivalent sensitivity values to models with similar misclassification costs for false negatives and false positives.

7.3.1.1 Interpretation of the Selected Models based on TS1

In the tree in Figure 7.2, the first split variable is SHHBd, which is the sum of the E-State indexes for hydrogen bond donors. This molecular descriptor is linked to the number of hydrogen bond donors highlighted in Lipinski's rule of five (Lipinski *et*

al., 1997). The cut off point for SHHBd is 7.81, which corresponds to roughly 3 or more hydrogen bond donor groups. Compounds with low hydrogen bonding donor ability (low SHHBd value) will have poor absorption if ABSQ, the sum of absolute values of atomic partial charges of the molecule, is high (node 5 in Figure 7.2). This indicates that molecules or compounds with electronegative or positive atoms (molecules containing heteroatoms) will be less absorbed through the intestine (Gasteiger and Marsili, 1980). This is in agreement with the hydrogen bond acceptor factor in Lipinski's rule of five. The compounds with low number of heteroatoms (ABSQ) will have high absorption unless they are highly acidic and have high ionization fraction at pH 1 (FiA1 > 0.139). It has been well cited that drugs that are unionised will pass better through the intestinal membrane (Lipinski *et al*., 1997, Pang, 2003).

The next important descriptor selected by C&RT for the partitioning of highly hydrogen bond donor compounds is VAMP HOMO, which is the energy of the highest occupied molecular orbital calculated by AM1 semi empirical method and has been used in previous QSAR models for bioavailability (Turner *et al*., 2003) and oral absorption (Agatonovic-Kustrin *et al*., 2001). According to this split, compounds with HOMO energy value ≤ -9.22 are all poorly-absorbed compounds (El-Henawy *et al*., 2013). These are highly polar molecules containing many hydrogen bonding groups (SHHBd) and few or no double bonds – e.g bisphosphonates and macrolides. The high HOMO energy group (Node 7) on the other hand, consists mainly of compounds of moderate absorption level (HIA of 40-60%) and, although marked as highly-absorbed, contains more of the poorly-absorbed compounds to be classified at the next level. These compounds are also of polar nature with many hydrogen bonding groups, but they also have planar areas in the molecule resulting from aromatic groups or other conjugated double bonds (hence high HOMO energy) (Eakins *et al*., 2011, Agatonovic-Kustrin *et al*., 2001). High HOMO energy compounds at Node 7 will have high absorption provided that they have dipole moment > 6.63. An inspection of these compounds at Node 9 shows that these are mainly natural or semi-synthetic compounds, e.g. contain a peptide or a sugar-like fragment in their structure. These compounds may be absorbed by carrier systems due to resemblance to natural metabolites. An example

of this is the compound oxytetracycline, which contains an aromatic system with many oxygen and nitrogen functional groups and is a known substrate of human organic anion transporters (Sai and Tsuji, 2004).

For Figure 7.3 (model 10), although all of the eight descriptors of descriptor set 1 were used as independent continuous variables in the C&RT analysis, not all of them were used to build the tree in Figure 7.3; in fact only six out of the eight were used, with SHBin7 and SsCH3 not being selected. Similar to model 9, SHHBd is the first split variable in this model. The compounds with high hydrogen bond capacity (according to SHHBd) with low absorption (node 3) has been partitioned again according to SpcPolarizability, which has replaced VAMP HOMO in the previous model (Figure 7.2). SpcPolarizability defines how readily the molecular charge distribution on a molecule is affected by external oscillating fields. It can also be described as the specific polarizability, which is equal to polarizability/volume. Compounds with low SpcPolarizability values have been divided into groups according to their SHBint2 values. SHBint2 is the sum of E-state indexes for hydrogen bonding groups of path length 2 (Wanchana *et al*., 2004) and is high in compounds like saquinavir and ceftriaxone which contain peptide bonds. If this value is high then compounds will be classed into the poor absorption class. Compounds with low SHHBd (node 2) have also been partitioned according to SHBint2, with chemicals containing a low number of hydrogen bonding groups with a bond distance of two showing high oral absorption probability (node 4). Compounds with high SHBint2 may still have high oral absorption if 'inertia moment 2 size' (a size related descriptor) has a low value and the ACDlogD5.5 (lipophilicity descriptor) value is high (node 11) and the ACD_Density (molecular density) value is small (node 12). Descriptors relating to molecular size have been inversely related to intestinal absorption, therefore the larger the molecule the lower the absorption (Varma *et al*., 2010). The relationship with logD (a measure of hydrophobicity at a specific pH) is in accordance with previous literature (Varma *et al*., 2010, Zakeri-Milani *et al*., 2006, Comer, 2003). ACD Density is the mass per unit volume of a molecule; density will be high for molecules containing many heteroatoms. Compounds with a high density will have low absorption which is also true according to this tree (Agatonovic-Kustrin *et al*., 2001).

7.3.2 C&RT Classification Analysis for TS2

C&RT classification analysis with misclassification costs was also carried out on TS2, the unbalanced dataset, to see if false positive error rates could be reduced. As there are a larger number of highly-absorbed compounds compared to poorly-absorbed compounds, the misclassification costs to reduce the number of false negatives does not need to be applied as the class distribution of TS2 already favours the decrease of false negatives. Therefore, misclassification costs are applied for reducing false positives only. The cost of 4 was applied to false positives (keeping the baseline cost of 1 for false negatives), as this was considered the most suitable number based on the class distribution of roughly 4:1 for high to low absorption compounds. The results of the C&RT classification analysis for TS2 are shown in Table 7.3.

Table 7.3. The results of C&RT classification analysis using different descriptor sets and misclassification cost ratios for TS2

Model	Cost FP:FN	Descriptor Set	Set	n Validation Set	SP × SE	SE	SP	CNMI
13	1:1	ALL	t	94	0.862	0.955	0.903	0.053
			v		0.400	0.880	0.455	0.170
14	1:1	3	t	102	0.704	0.973	0.724	0.064
			v		0.445	0.954	0.467	0.118
15	1:1	4	t	102	0.620	0.982	0.632	0.070
			v		0.451	**0.966**	0.467	0.108
16	4:1	ALL	t	94	0.861	0.873	0.986	0.033
			v		**0.660**	0.807	**0.818**	**0.070**
17	4:1	3	t	102	0.879	0.890	0.987	0.028
			v		0.653	0.816	0.800	0.077
18	4:1	4	t	102	0.855	0.890	0.961	0.033
			v		0.517	0.862	0.600	0.099

FP = False positive; FN = False negative; t-training; v-validation; Sensitivity is equivalent to the number of correctly classified highly-absorbed compounds and is calculated using SE=(TP/(TP+FN)); Specificity is equivalent to the number of correctly classified poorly-absorbed compounds and is calculated using SP=(TN/(TN+FP)); TP-true positive; FN-False negative; TN-true negative; FP-false positive; Overall Accuracy of the models was calculated by multiplying Specificity by Sensitivity (SP x SE); n, validation is the number of validation set compounds that was predicted by the model CNMI = Cost normalised misclassification index

Table 7.3 shows that when all descriptors were available to C&RT analysis, the best results were achieved when applying misclassification costs to reduce false positives (comparing model 13 and 16). As expected, specificity increases and misclassification error rate decreases when higher misclassification costs were applied to false positives. By applying misclassification costs to increase specificity, the sensitivity of the model will decrease (Table 7.3). Figure 7.4 shows the best model when all descriptors were supplied and the significant descriptors were selected by C&RT analysis (model 16).

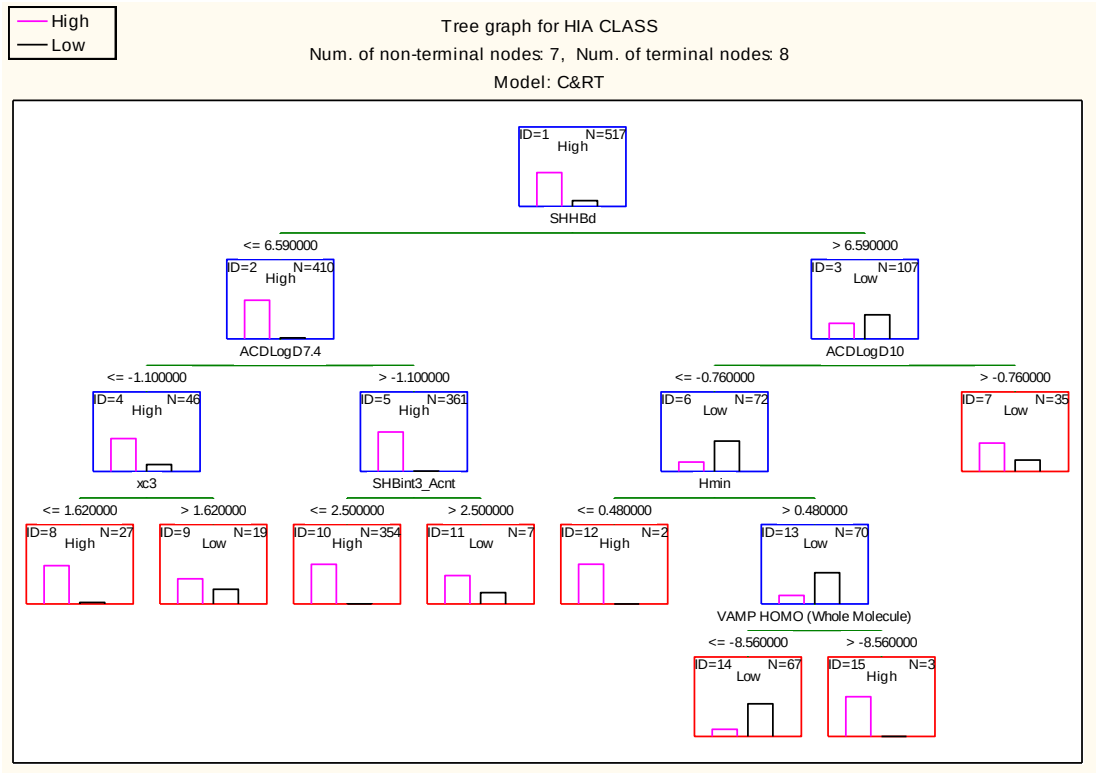

Figure 7.4. Tree graph for the best C&RT analysis using TS2 using all descriptors with misclassification costs applied to reduce false positives (Model 16)

As with the TS1 training set, C&RT analysis was also carried out using the pre-selected molecular descriptors from the previous chapter 6. Table 7.3 shows that the best pre-selected descriptor set was descriptor set 3 (model 17) when considering SP × SE. The classification tree model 17 is shown in Figure 7.5. With misclassification costs to reduce false positives, the tree had the highest specificity (0.800) and also the lowest CNMI (0.077). Depending on the use of the model, if the reduction of

false positives (increase of specificity) is the intention, then using misclassification costs will increase the specificity for descriptor set 3 from 0.467 to 0.800; however, sensitivity decreases from 0.954 to 0.816.

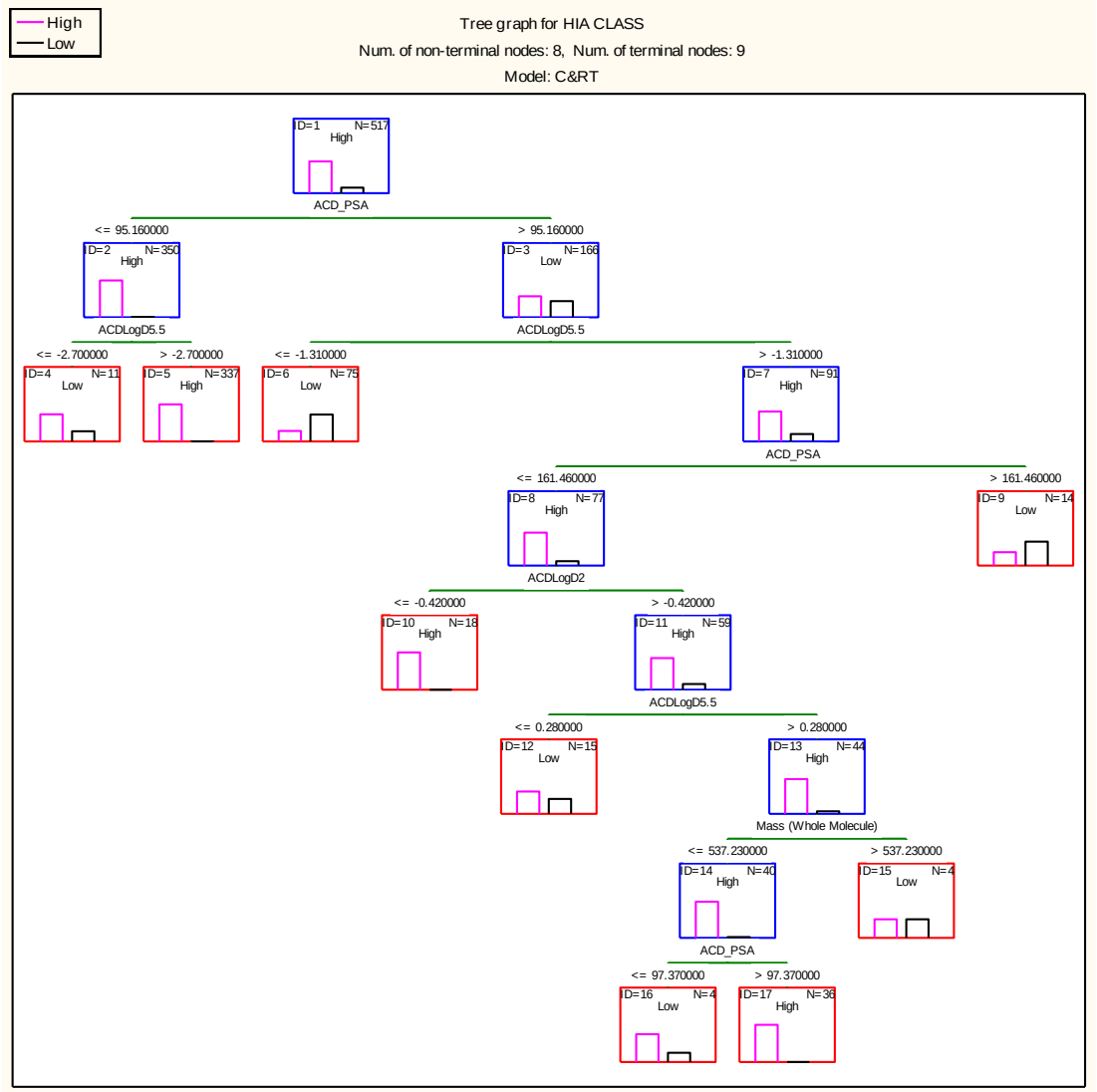

Figure 7.5. Tree graph for C&RT analysis using TS2 with misclassification costs applied to false positives (FP:FN 4:1) using descriptor set 3 (Model 17)

7.3.2.1 Interpretation of Selected Model based on TS2

Figure 7.4 shows the selected tree when C&RT analysis selected the descriptors from all the supplied descriptors (model 16). Similar to models 9 and 10 obtained using TS1, this tree involves the hydrogen bond donor descriptor, SHHBd, as the first variable. Compounds with high SHHBd values are more likely to have poor oral absorption, especially if they are hydrophilic with ACDLogD10 below -0.76; unless

their Hmin value is lower than 0.48. A high number of potential H-bond formations is detrimental to high oral absorption, which is cited in the literature (Lipinski, 2000, Lipinski *et al*., 1997). Hmin is the minimum hydrogen electrotopological-state value for all atoms in the drug molecule and shows the nature of the hydrogen atoms attached to the skeleton of the drug molecule and whether they are hydrogen bond donors (Wang *et al*., 2008). Otherwise, if the Hmin value is higher than 0.48, compounds with VAMP HOMO higher than -8.56 may still have high oral absorption, but the large majority of compounds have a lower HOMO energy value and therefore will be expected to be poorly-absorbed through the gastrointestinal system (node 14). On the left hand side of the tree (node 2), for compounds with low hydrogen bond donor ability (SHHBd ≤ 6.59), oral absorption is expected to be high, unless ACDlogD7.4 is low and xc3 is high (node 9). The descriptor xc3 is the third order cluster chi connectivity index. This Chi index encodes the extent of branching of the molecule and in this tree, it indicates that branched molecules (of hydrophilic nature) have poor oral absorption (Hall and Kier, 1991). It must be noted in Figure 7.4 in nodes 9 and 11, for example, that the effect of misclassification costs is altering the final terminal class node, showing the misclassification costs applied to reduce false positives is working. Moreover, for higher ACDlog7.4 compounds (> -1.10) oral absorption would be poor if they have a high number of hydrogen bonding groups with a bond distance of three (SHBint3_Acnt). It is interesting to note that the nodes in the tree after the first split using SHHBd were both divided by logD but measured at different pH values. LogD at different pH values is affected by the ionization of the compound and is related to the compound's pKa. For example, for logD10, which means the distribution coefficient at pH10, any basic compounds at pH10 will be unionized, therefore will have higher logD10 values than acidic compounds which will remain ionized due to the higher pH, and in the case of intestinal absorption will then be not absorbed. This indicates that the pH-dependent lipophilicity measure (logD) at different pH values is important in distinguishing between high and low absorption for acidic and basic compounds as well as characterizing the lipophilicity.

In Figure 7.5, all the molecular descriptors have been described previously apart from PSA. This descriptor along with lipophilicity has been described as an

influential molecular descriptor in predicting passive intestinal absorption and has been described on page 97 (van de Waterbeemd and Gifford, 2003, Palm *et al*., 1997, Palm *et al*., 1996). Mass has also been used in this tree and in accordance with Lipinski's rule of five, but only with a slightly different cut off point of 537.23 Da (Lipinski *et al*., 1997, Yang *et al*., 2012).

7.4 Discussion

The aim of this chapter was to create classification models that could classify compounds into highly and poorly-absorbed classes using two techniques to overcome datasets with unbalanced class distribution: under sampling and misclassification costs. Firstly, under-sampling was used to create a balanced training set to build models to compare with an unbalanced training set using decision trees. Secondly, higher misclassification costs were applied to the balanced and unbalanced datasets to see the effect the higher misclassification costs have on model accuracy and overcoming the dataset with unbalanced class distribution. A comparison of models in this chapter as well as previously published oral absorption models is discussed.

7.4.1 Comparison of Models

As stated in the previous chapter, a direct comparison between models built from the two different training sets is not a fair comparison due to the different class distributions of TS1, the balanced set, and TS2, the set biased toward highly-absorbed compounds. Nevertheless, it can be seen that TS1 in the majority of cases leads to higher specificity when misclassification costs are equal for FN and FP. TS2 gave higher sensitivity in all cases, which is expected due to the bias of the training set toward highly-absorbed compounds. It has been cited that Lipinski's rule of five can give rise to false positives and could be a possible explanation for the lower specificity of this model even with a balanced training set (Lipinski *et al*., 1997, Andrews *et al*., 2000, Zhu *et al*., 2011). When misclassification costs are applied to either TS1 or TS2 to reduce false positives, specificity improves for both training sets. Moreover, it can be seen in Table 7.3 that the use of 4:1 misclassification costs for FP:FN leads to improved models for TS2 with better SP × SE values. This

finding shows that using misclassification costs can overcome a dataset bias by increasing specificity.

Additionally in this chapter, I compared the effect of allowing the software to pick the most significant descriptors from all 215 descriptors used or from a smaller subset of descriptors previously selected as significant by stepwise regression analysis or those related to Lipinski's rule of five. For the balanced training set, TS1, Table 7.2 shows that model 10 achieved the lowest value CNMI of 0.165 and the highest SP × SE of 0.686 using descriptor set 1. The next best CNMI was again achieved by descriptor set 1 with a value of 0.188 (model 6); this model also obtained the highest specificity of 0.925 when misclassification costs were applied to reduce false positives. From Table 7.2 it is interesting to see that in several cases the CNMI values are higher for C&RT models using all descriptors compared with those models using smaller descriptor sets selected by feature selection techniques, meaning that there are more errors when allowing the C&RT analysis to pick significant descriptors from the 215 available. This could show that using linear stepwise regression to select a smaller subset of molecular descriptors significantly relevant to intestinal absorption beforehand can be advantageous as often models are produced with fewer misclassifications. The most accurate models for TS2 (Table 7.3) are models 16 and then 17, which were developed using all descriptors or descriptor set 3. The fact that using all descriptors works well for TS2, but for TS1 a prior descriptor selection is best, suggests that C&RT can be an efficient descriptor selection method when a larger dataset is used (517 vs 94 compounds in TS2 and TS1, respectively).

In this chapter, TS1 contained 94 compounds with a validation set containing 89 compounds and was used as this mirrors the balanced data distribution of the training set. By balancing the validation set too, it gives a fair representation of the models' predictive performance. As an additional test, the predictive performance of the models was investigated for a new validation set containing all the compounds not used in the training set, plus the 89 compounds already in the validation set. It must be noted that the additional validation set compounds are all highly-absorbed with the exception of two compounds. Therefore this validation set is heavily biased. The

results of this work can be found in Table 7.4. According to Table 7.4 the best models according to SP × SE were those using descriptor set 1 (models 2, 6, and 10), which corresponds to the results seen earlier for the smaller balanced validation set (Table 7.4 model 10). The performance on the larger validation can help confirm how well the models using the balanced training set perform.

Table 7.4. The validation results of C&RT classification models obtained using TS1 for all the remaining compounds not used in training

Model	Cost FP:FN	Descriptor Set	n validation set	SP x SE	SE	SP	CNMI
1	1:1	ALL	496	0.647	0.869	0.745	0.143
2		1	521	**0.730**	0.852	0.857	0.148
3		2	521	0.728	0.887	0.820	0.119
4		4	521	0.341	0.898	0.380	0.152
5	2:1	ALL	496	0.510	0.800	0.638	0.131
6		1	521	0.712	0.775	**0.918**	0.115
7		2	521	0.685	0.856	0.800	0.089
8		4	521	0.654	**0.909**	0.720	**0.072**
9	1:2	ALL	496	0.673	0.855	0.787	0.258
10		1	521	0.701	0.881	0.796	0.214
11		2	521	0.657	0.887	0.740	0.208
12		4	521	0.497	0.887	0.560	0.224

FP = False positive; FN = False negative; t-training; v-validation; Sensitivity is equivalent to the number of correctly classified highly-absorbed compounds and is calculated using SE=(TP/(TP+FN)); Specificity is equivalent to the number of correctly classified poorly-absorbed compounds and is calculated using SP=(TN/(TN+FP)); TP-true positive; FN-False negative; TN-true negative; FP-false positive; Overall Accuracy of the models was calculated by multiplying Specificity by Sensitivity (SP x SE); n, validation is the number of validation set compounds that was predicted by the model CNMI = Cost normalised misclassification index

7.4.2 Discussion of the Related Literature

As specified in the previous chapter, there are summary tables in the literature that detail the accuracy, specificity and sensitivity of classification work carried out by previous studies (Hou *et al*., 2006, Suenderhauf *et al*., 2011, Talevi *et al*., 2011). As previously stated, the direct comparison of the models presented in this chapter with the models in the literature is a very difficult task. In spite of this the comparison of those models using similar classification techniques can be discussed in the context of the biased dataset problem.

The number of compounds in the datasets in the literature should be considered when assessing the models' performances. Small datasets may achieve high predictive accuracy within the chemical space of the dataset, however this can lead to a lack of generalization to new chemical compounds which are likely to be outside the domain applicability of such models. In the study by Niwa *et al* (2003), using 67 compounds achieved 100% correct classification for the training set, however this dropped to 80% for the external prediction set of 12 compounds. It must be highlighted that the main misclassification in Niwa *et al*'s model was for the poorly-absorbed compounds, which were represented inadequately in Niwa's dataset (Niwa, 2003). As a result, the overall accuracy of Niwa *et al*'s model as calculated using my accuracy measurement (SP × SE) yields a value of 0.667. This is a reoccurring problem with the other datasets in the literature that I considered (Wessel *et al*., 1998, Zhao *et al*., 2002, Klopman *et al*., 2002). Poorly-absorbed compounds are predicted better using my models due to the larger representation of this class in my TS1 training set and/or the use of varying misclassification costs.

An interesting study carried out by Bai *et al* using C&RT analysis found that using an even distribution of compounds with low, moderate and high absorption resulted in similar prediction ability for the test set using an unbalanced dataset. They summarised that the training set distribution does not affect how the C&RT algorithm performs (Bai *et al*., 2004); this is contrary to the results obtained in this chapter. C&RT can deal with skewed datasets to some extent, but this chapter has highlighted that using an unbalanced dataset gives rise to poorer predictive accuracy for the under-represented class using the C&RT method. However, the use of misclassification costs and under sampling can result in models with higher predictive accuracy in general, based on the work in this chapter.

Additionally, Deconinck *et al* (2005) carried out C&RT analysis using Splus software using the 141 compounds from Zhao *et al* (2001). C&RT models were built using different subsets of 2D and 3D molecular descriptors and validated on a small validation set of 27 compounds. All of the 27 compounds were highly-absorbed; hence, the accuracy measure is really the sensitivity value of 85%. With no poorly-

absorbed compounds in the validation set, the real predictive ability of the model for both classes cannot be confirmed using a suitable external validation set. In spite of this, this study highlighted that C&RT is a suitable technique to select important molecular descriptors for oral absorption (Deconinck *et al*., 2005).

As stated in the previous chapter, in most studies the accuracy and sensitivity results are higher than the specificity values, due to the under-representation of poorly-absorbed compounds (Deconinck *et al*., 2005, Niwa, 2003). The exceptions to this were obtained by Hou *et al* (2007), who obtained higher specificity than sensitivity using a validation set of 98 compounds. However, their result was achieved when only five of the 98 compounds in their validation set were poorly-absorbed. The two papers utilise recursive partitioning (RP), a type of decision tree method, and SVM for the prediction of high and low absorption using a threshold of 30% to define the boundary between the two classes. On closer inspection of the study using SVM, it was highlighted that cost sensitive learning was applied. In other words, higher misclassification costs were applied to the minority class. Without higher misclassification costs applied to the SVM models, the specificity was lower than the sensitivity and overall accuracy, indicating that SVM's classification can be influenced by the unbalanced dataset.

Lipinski's rule of five is a qualitative rule-based model which has been explained in previous chapters. This rule has been criticised for having a high rate of false positives (Zhu *et al*., 2011, Suenderhauf *et al*., 2011). With this work, descriptors describing Lipinski's rule of five plus the number of rotatable bonds allowed a qualitative evaluation of Lipinski's rule of five via C&RT analysis. Using Lipinski's rule of five in its original form (if 2 or more of the 5 rules were violated, this indicates poor absorption), specificity was 0.425 and 0.400 for the validation sets of TS1 and TS2 respectively. By incorporating these descriptors (descriptor set 4) in C&RT, upon using higher misclassification costs to reduce false positives, the specificity was 0.750 for TS1 and 0.600 for TS2. Using misclassification costs to reduce false positives, an improvement to Lipinski's rules was made possible.

The descriptors selected in the models can be interpreted according to the known mechanisms involved in the absorption process. Table 7.5 gives a summary of all the molecular descriptors used in the selected models, 9, 10, 16 and 17. The most common molecular descriptors used in the best models were descriptors of hydrogen bonding (such as SHHBd, SHBint2); log D at various pH values, which is related to lipophilicity and acid/base property; ACD_Density, which is related to the number of heteroatoms in the molecules; and PSA, which has been cited as a molecular descriptor relating to polarity and size (Hou *et al*., 2007c, Wegner *et al*., 2004). Other important molecular attributes are size-related parameters. These are in agreement with the literature indicating that the molecular descriptors important to intestinal absorption are those related to lipophilicity, hydrogen bonding, polarity, ionization, and size (Yang *et al*., 2012, Wegner *et al*., 2004). Overall, the molecular descriptors utilised in the best models in this chapter, no matter what training set used, are shown in the literature to be important for intestinal absorption (Deconinck *et al*., 2005)

Table 7.5. Molecular Descriptors Used in the Selected Models (9, 10, 16 and 17)

Type of descriptor	Name of descriptor	Number of selected models containing the descriptor
Hydrogen bonding	SHHBd	3
	ABSQ	1
	ACD_Density	2
	SHBint2	1*
	SHBint3_Acnt	1
	Hmin	1
	ACD_PSA	1*
Lipophilicity	ACD logD 5.5	2*
	ACD logD 7.4	1
	ACD logD 10	1
	ACD logD 2	1
Size	xvch7	1
	Inertia moment 2 (size)	1
	xc3	1
	Mass	1
Polarity/ polarizability	VAMP HOMO	2
	Total dipole moment	1
	Spc polarizability	1
Acidity	FiA1	1

* occurred more than once in a single tree model

7.5 Conclusion

Class imbalance occurs frequently in QSAR and drug discovery datasets (Tropsha, 2010). This could be for a number of reasons; however, in this context it is due to lack of publicly available data for the minority class, poorly-moderately absorbed compounds, in the literature. The aim of this work was to improve the class prediction of poorly-absorbed compounds by the use of varying misclassification costs in C&RT analysis. This was analysed using two training sets, the one selected by under-sampling the majority class (TS1), or the training set selected randomly and hence biased towards highly-absorbed compounds (TS2). The comparison between descriptor selection by C&RT and pre-selecting a small subset of molecular descriptors using statistical techniques or rule-based models was also considered.

Under-sampling the majority class to create a balanced training set produced models that had high predictive power for the prediction of poorly-absorbed compounds. As expected, the randomly selected training set (TS2) had high predictive power for highly-absorbed compounds with high sensitivity values, but this was accompanied by low specificity values. This conclusion conforms to the previous work using regression and discriminant analysis classification in the previous chapter (Ghafourian *et al*., 2012).

The use of misclassification costs led to improvements in predictive accuracy. Even though there is no general consensus to reduce false positives or false negatives from the literature, this work shows that misclassification costs can be applied to reduce false positives or false negatives. Other considerations such as poor solubility and carrier-mediated transport systems can play a part in misclassification error rates in the models (Klopman *et al*., 2002). For the unbalanced training set containing the majority high absorption class, applying higher costs for the misclassification of false positives improved specificity in all cases. The unbalanced dataset can be utilised without removing compounds as an advantage for improved sensitivity as it will already be biased towards high absorption compounds. Therefore, varying ratios of misclassification costs can be used as a vital and effective tool to overcome class unbalance, which is a recurring problem in drug discovery datasets.

The comparison between using all descriptors for the C&RT method or to use a smaller subset of molecular descriptors suggests that the descriptors selected by stepwise linear regression may achieve better prediction. However, this cannot be generalized, and descriptor selection by C&RT may work just as well when a large training set is used, e.g. TS2. This warrants further investigation.

In conclusion, reasonably interpretable, user friendly C&RT models that can be easily understood and utilised for specific purposes have been obtained by using two strategies, under-sampling the majority class of the training set and misclassification costs, to overcome class imbalance of oral absorption datasets.

8 Pre-processing Feature Selection for Improved C&RT Models for Oral Absorption

8.1 Introduction

In order to construct a QSAR model, molecular descriptors are calculated in order to derive mathematical relationships between the chemical structure and the biological activity of compounds. There are a wide variety of molecular descriptors to choose from; therefore feature selection techniques can be used to identify the best ones.

As previously stated in the introduction, feature selection techniques selectively minimize the number of molecular descriptors used to accurately describe property of interest, in this case %HIA class (Wong and Burkowski, 2011). Therefore, feature selection can increase interpretability, predictive accuracy and reduce over fitting for subsequent models (Goodarzi *et al*., 2012). Feature selection can be categorized into the two broad categories of data pre-processing or embedded methods. Data pre-processing feature selection involves reduction of the number of molecular descriptors before model building. On the other hand, embedded methods incorporate feature selection into the training and building of the final model (Goodarzi *et al*., 2012, Saeys *et al*., 2007).

Based on the previous chapters where descriptor subsets were used to create models for oral absorption, this chapter investigates five pre-processing filter feature selection techniques for selecting subsets of molecular descriptors. The comparison of these different feature selection techniques is anticipated to give an idea of the relative abilities of the different techniques based on their prediction ability on the validation set. Furthermore, I compare two broad approaches for feature selection: (1) a "two-stage" feature selection procedure, where in the first stage a pre-processing feature selection method selects a subset of descriptors, and in the second stage classification and regression trees (C&RT), which is itself an embedded feature selection method, selects a subset of the descriptors selected by the filter technique to build a decision tree; (2) a "one-stage" approach where C&RT is used as the only feature selection technique, without using data pre-processing feature selection methods (Figure 8.1).

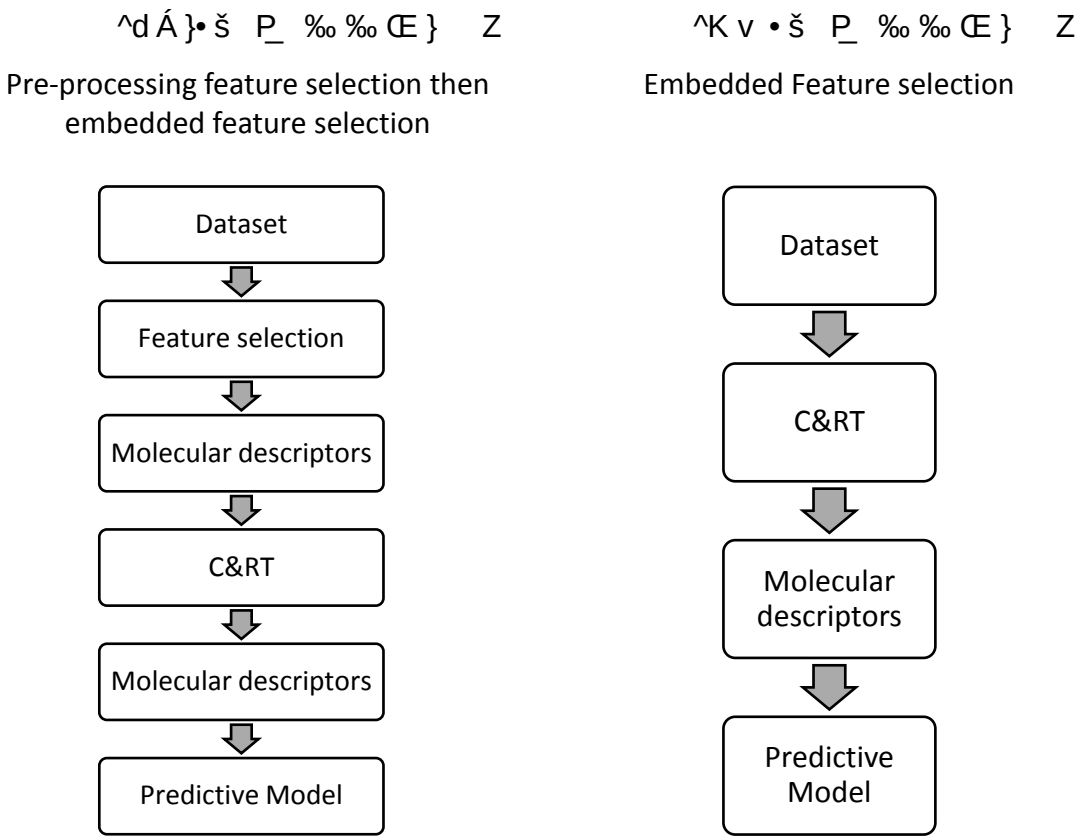

Figure 8.1. A comparison of pre-processing and embedded feature selection processes for model building

A comparison between these two approaches in Figure 8.1 could indicate the usefulness of pre-processing feature selection for C&RT analysis. Additionally, this work utilizes misclassification costs in model building to overcome the problem of imbalanced class datasets as used in Chapter 7. Therefore, this section offers an investigation of feature selection techniques which reduce the number of molecular descriptors, increasing interpretability of resulting models, and combined with this the use of misclassification costs in model development to increase a model's predictive accuracy when analysing an imbalanced class dataset. Therefore this work offers a novel combination of pre-processing feature selection combined with misclassification costs to develop models for imbalanced class oral absorption datasets.

8.2 Methods

8.2.1 Dataset

Dataset 2 was utilised for the prediction of oral absorption as defined in the 'Dataset and Methods' section 5.1.2 of this thesis.

8.2.2 Training Sets and Validation Sets

There were three sets of compounds used in this work: a training set, to train the models, a parameter optimisation set, for method optimisation and finally an external validation set, to assess the predictive ability of the models produced with an unseen validation set. The training set and parameter optimisation set (internal validation set) were assigned based on previous chapters using dataset 1. The additional data were used as an external validation set and when combined with the compounds from dataset 1 resulted in dataset 2.

All compound sets (training, parameter optimisation and validation sets) had similar data distributions to create a fairer more controlled validation of the models. The exact number of compounds in the training, parameter optimisation and validation set are shown in Table 8.1.

Table 8.1. Numbers of compounds in the training, parameter optimisation and validation sets

Dataset	**Number of compounds (n)**
Training set	534
Parameter optimisation set	107
Validation set	48

8.2.3 Model Development

8.2.3.1 Molecular Descriptors

A total of 204 descriptors were generated in this study using a variety of different software including TSAR 3D (Accelrys Inc.), MDL QSAR (Symyx Inc.), MOE

(Chemical Computing Group Inc.) v2010.10 and Advanced Chemistry Development ACD Labs/LogD Suite v 12. From the 204 molecular descriptors, 203 were continuous and only one was categorical.

In this work it was noted that compounds that contained a permanent quaternary ammonium ion had more missing descriptor values than other compounds in the data set. Therefore, an indicator variable that described the permanent positive nitrogen (YES/NO) was calculated using the MOE software (v2012). Molecular descriptors that are difficult to compute and result in many missing values may not be suitable to be used in resulting models as the molecular descriptors may not be able to be calculated for new compounds, leading to poor performance of the model for classification of these compounds. Therefore, all molecular descriptors that had 10 or more missing values based on preliminary work were removed and therefore a final number of 204 descriptors were available for feature selection techniques.

8.2.3.2 Feature Selection

Feature selection methods were used in the pre-processing step to reduce the number of molecular descriptors to smaller subsets that accurately predict the HIA Class. The software used for feature selection was STATISTICA v11 and WEKA v 3.6 (Hall *et al*., 2009). The pre-processing feature selection techniques used to select molecular descriptors for the classification models of oral absorption are shown in Table 8.2 and are further described in the general methods section.

Table 8.2. Pre-processing feature selection methods utilised in this chapter

	Feature selection method	**Acronym used in this chapter**	**Software used**
1	Predictor importance using random forest	RF	STATISTICA
2	Predictor importance using random forest with higher misclassification costs for false positives*	RF (MC)	STATISTICA
3	Chi-square	CS	STATISTICA
4	Information gain ratio	IGR	WEKA
5	Greedy stepwise	GRD	WEKA
6	Genetic search	GEN	WEKA

*Higher misclassification costs applied 4:1 False Positive: False Negative

The molecular descriptors selected by the pre-processing feature selection techniques in Table 8.2 were used as input by C&RT which then performed further (embedded)

feature selection (Figure 8.1) in order to build the resulting models. The training set was used by all methods; however, for the filter methods CS, IGR, GRD, and GEN, the parameter optimization set was combined with the training set to carry out feature selection using these techniques. For random forest and C&RT (embedded feature selection) the training set was used to train the model, and separately the parameter optimization set was used to assess the optimisation of the model parameters based on the training set. The top 20 molecular descriptors for methods RF, CS, and IGR were selected based on the highest values of the descriptor scoring function. Other numbers of selected molecular descriptors were tried; however, based on the C&RT analysis results on the parameter optimization set, the top 20 descriptors gave the highest classification accuracy and this was selected.

8.2.3.3 QSAR Modelling Techniques

The categorical prediction of HIA class was carried out using C&RT analysis in STATISTICA v 11 (StatSoft Ltd). According to observed %HIA values in the dataset, compounds were placed into either the 'High' class if %HIA was equal to or greater than 50% or the 'Low' class, if %HIA was less than 50%.

For this work, HIA Class was set as the dependent categorical variable and either all 203 molecular descriptors or a subset of these selected by various feature selection methods in Table 8.2 were selected as continuous independent variables. The analyses also included one categorical independent variable, N+ group, the indicator variable for presence or absence of quaternary ammonium. If there were any trees with only one compound in the terminal nodes, manual pruning was carried out to prevent over fitting. In C&RT analysis, the stopping factor was the minimum number of compounds for splitting of 30 which was selected based on preliminary experiments. All other settings used were default settings defined by the software.

C&RT carries out embedded feature selection. Therefore the use of feature selection methods in a pre-processing phase, before inputting the descriptor subset into C&RT can be investigated. By carrying out data pre-processing feature selection the method can avoid C&RT's drawback of 'data fragmentation'. In other words, as the decision tree is built and compounds split into smaller nodes there are fewer compounds to split; therefore, the selection of descriptors in that local node becomes less

statistically reliable. Whereas if a statistically significant subset of molecular descriptors relating to HIA class has already been selected via pre-processing feature selection and then are given to C&RT, the selection of molecular descriptors will be more reliable.

8.2.3.4 Misclassification Costs

For this chapter, either equal misclassification costs were assigned or a misclassification cost ratio of 4:1 was applied to false positive:false negatives based on the dataset distribution, as in the previous chapter. This chapter compared the use of the same costing with higher costing to reduce false positives as well as comparison of the pre-processing and embedded feature selection approaches.

8.3 Results

A full list of molecular descriptors selected by each of the feature selection methods can be found in Appendix 2, Tables A2.1 and A2.2. As GRD and GEN are not ranking feature selection methods, the number of descriptors picked by these methods will depend on the technique and the dataset. GRD selected a total of 21 descriptors and GEN selected 64. Tables 8.3 and 8.4 show the predictive performance measures of the classification trees using different sets of molecular descriptors selected by the feature selection methods. In Table 8.3 equal misclassification costs have been applied to false positives and false negatives for C&RT analysis, while in Table 8.4 the ratio of misclassification costs is 4:1 for false positives: false negatives. In Table 8.3 and 8.4 the best models are those that have the highest SE, SP and SP x SE measures and the lowest CNMI. These have been highlighted in **bold** for the training (t), parameter optimisation (po) and validation (v) sets. For the random forest feature selection method there was an option to apply misclassification costs. Therefore the descriptor sets selected by RF with equal (models 1 and 8) and higher misclassification costs applied to false positives (models 2 and 9) were used and also compared. All the C&RT decision trees from Tables 8.3 and 8.4 can be found on the accompanying disk with this thesis.

Table 8.3. The results of C&RT classification analysis using different feature selection methods with equal misclassification costs applied to the C&RT algorithm

Model	Feature selection Method	dataset	n	SP x SE	SE	SP	CNMI
1	RF	t	531	0.848	0.950	0.892	0.060
		po	107	0.709	0.930	0.762	0.103
		v	47	0.363	0.816	0.444	0.255
2*	RF (MC)	t	531	**0.884**	0.945	**0.935**	0.056
		po	107	**0.757**	0.884	**0.857**	0.121
		v	47	**0.453**	0.816	**0.556**	0.234
3	CS	t	531	0.777	0.963	0.806	0.064
		po	107	0.576	0.930	0.619	0.131
		v	47	0.187	0.842	0.222	0.277
4	IGR	t	531	0.800	**0.979**	0.817	0.049
		po	107	0.664	0.930	0.714	0.112
		v	47	0.398	**0.895**	0.444	**0.191**
5	GRD	t	531	0.803	0.970	0.828	0.055
		po	107	0.628	**0.942**	0.667	0.112
		v	47	0.351	0.789	0.444	0.277
6	GEN	t	531	0.839	0.975	0.860	**0.045**
		po	107	0.673	**0.942**	0.714	0.103
		v	47	0.398	**0.895**	0.444	**0.191**
7	C&RT	t	531	0.784	0.959	0.817	0.066
		po	105	0.694	**0.942**	0.737	**0.095**
		v	47	0.281	0.842	0.333	0.255

t-training; po, parameter optimisation; v-validation; Sensitivity is equivalent to the number of correctly classified highly-absorbed compounds and is calculated using SE=(TP/(TP+FN)); Specificity is equivalent to the number of correctly classified poorly-absorbed compounds and is calculated using SP=(TN/(TN+FP)); TP-true positive; FN-False negative; TN-true negative; FP-false positive; Overall Accuracy of the models was calculated by multiplying Specificity by Sensitivity (SP x SE); n, is the number of compounds that was predicted by the model for the training, parameter optimisation and validation set; CNMI = Cost normalised misclassification index; * misclassification costs applied to feature selection method

Comparing models built with equal misclassification costs (Table 8.3); the best overall model to choose would be model 2. This model has the highest SP x SE, plus the highest specificity values for the training, parameter optimisation and validation sets. However, this model does not achieve the highest SE values, with SE = 0.945, 0.884 and 0.816 for the training, parameter optimisation and validation sets respectively. All other models have better SE than model 2 for the three data subsets;

apart from model 1, which has the same SE for the validation set, and model 5 (GRD), with a lower SE of 0.789. If the aim of the model was to achieve the best sensitivity then model 6, using genetic search feature selection, would be the best model to use, as it achieved the best sensitivity for the parameter optimisation and the highest SE for the training set amongst the three selected models above, along with the lowest CNMI for the training set. Model 2 was able to classify correctly all the permanent ammonium-containing compounds used in the training and parameter optimisation set, and this resulted in the correct prediction of a permanent ammonium containing compounds in the validation set. The classification tree using the molecular descriptors from this model is shown in Figure 8.2.

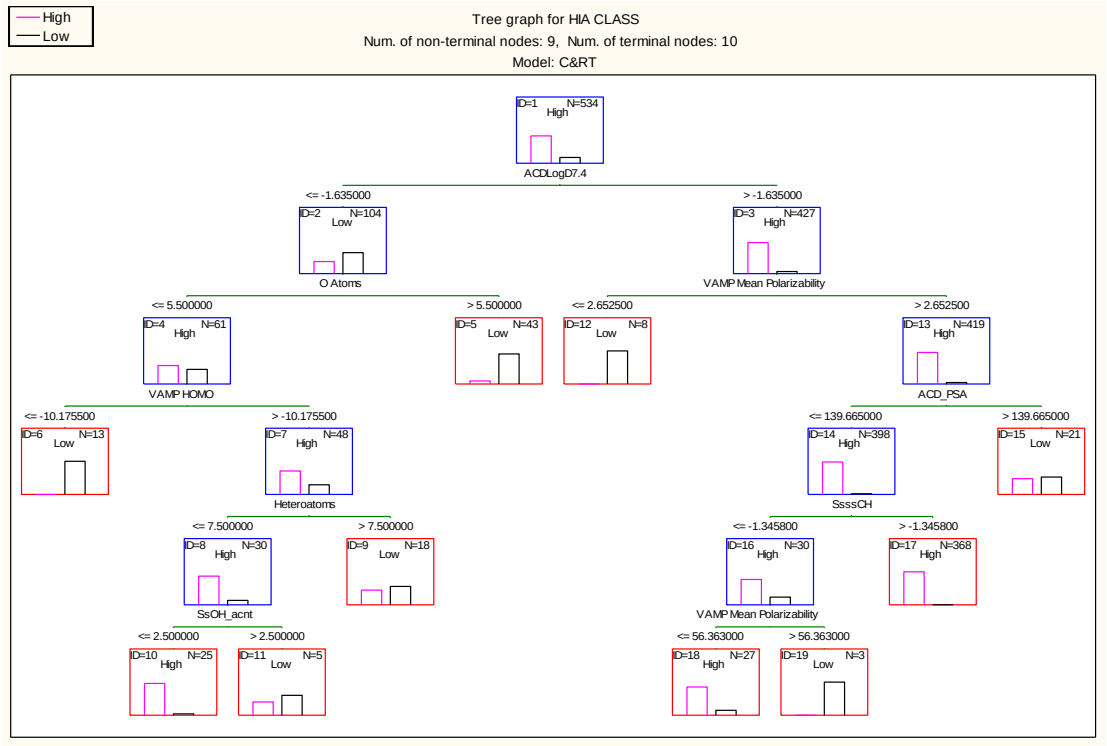

Figure 8.2. Tree graph for C&RT analysis using random forest predictor importance as feature selection method with equal misclassification costs applied to pre-processing C&RT (Model 2 in Table 8.3)

Table 8.4 shows the predictive performance measures from the classification trees using different sets of molecular descriptors selected by the feature selection methods when the ratio of misclassification costs is 4:1 for false positives: false negatives for C&RT analysis.

Table 8.4. The results of C&RT classification analysis using different feature selection methods with higher misclassification costs applied to false positives to the C&RT algorithm (misclassification cost ratio of FP: FN = 4:1)

Model	Feature selection Method	dataset	n	SP x SE	SE	SP	CNMI
8	RF	t	531	0.887	0.927	0.957	0.026
		po	107	0.725	0.895	0.810	0.068
		v	47	0.675	**0.868**	0.778	0.081
9*	RF (MC)	t	531	0.879	0.909	0.968	0.028
		po	107	**0.738**	0.860	**0.857**	**0.066**
		v	47	0.635	0.816	0.778	0.093
10	CS	t	531	0.838	0.906	0.925	0.037
		po	107	0.687	0.849	0.810	0.079
		v	47	0.544	0.816	0.667	0.118
11	IGR	t	531	0.853	0.934	0.914	0.033
		po	107	0.673	0.884	0.762	0.082
		v	47	0.544	0.816	0.667	0.118
12	GRD	t	528	0.892	**0.943**	0.946	0.025
		po	106	0.654	0.872	0.750	0.085
		v	47	**0.725**	0.816	**0.889**	**0.068**
13	GEN	t	531	0.885	0.895	**0.989**	0.027
		po	107	0.640	0.895	0.714	0.090
		v	47	0.614	0.789	0.778	0.099
14	C&RT	t	531	**0.911**	0.932	0.978	0.020
		po	107	0.726	**0.907**	0.800	**0.066**
		v	47	0.544	0.816	0.667	0.118

t-training; po, parameter optimisation; v-validation; Sensitivity is equivalent to the number of correctly classified highly-absorbed compounds and is calculated using SE=(TP/(TP+FN)); Specificity is equivalent to the number of correctly classified poorly-absorbed compounds and is calculated using SP=(TN/(TN+FP)); TP-true positive; FN-False negative; TN-true negative; FP-false positive; Overall Accuracy of the models was calculated by multiplying Specificity by Sensitivity (SP x SE); n, is the number of compounds that was predicted by the model for the training, parameter optimisation and validation set; CNMI = Cost normalised misclassification index; * misclassification costs applied to feature selection method

For Table 8.4, based on the SP x SE for the external validation set, the best model is model 12 with a SP x SE value of 0.725; but this model also had one of the lowest SP x SE for the po set (0.654), which has a higher number of chemicals compared to the external validation set. In comparison, models 8 and 9 achieved higher SP x SE of 0.725 and 0.738 respectively, where the po set was not used for molecular descriptor selection and hence it was also an external set. From models 8 and 9,

model 9 had a similar balance of high estimation of SP and SE compared to model 8, which was slightly worse at predicting poorly-absorbed compounds for the po set. What was interesting to note about model 9 was the feature selection method using predictor importance from random forest, which allowed misclassification costs to be applied at the feature selection level. Then the resulting C&RT model (with misclassification costs) achieved high predictive accuracy for the unseen validation set as well as training and parameter optimisation sets. The C&RT tree for model 9 is shown in Figure 8.3.

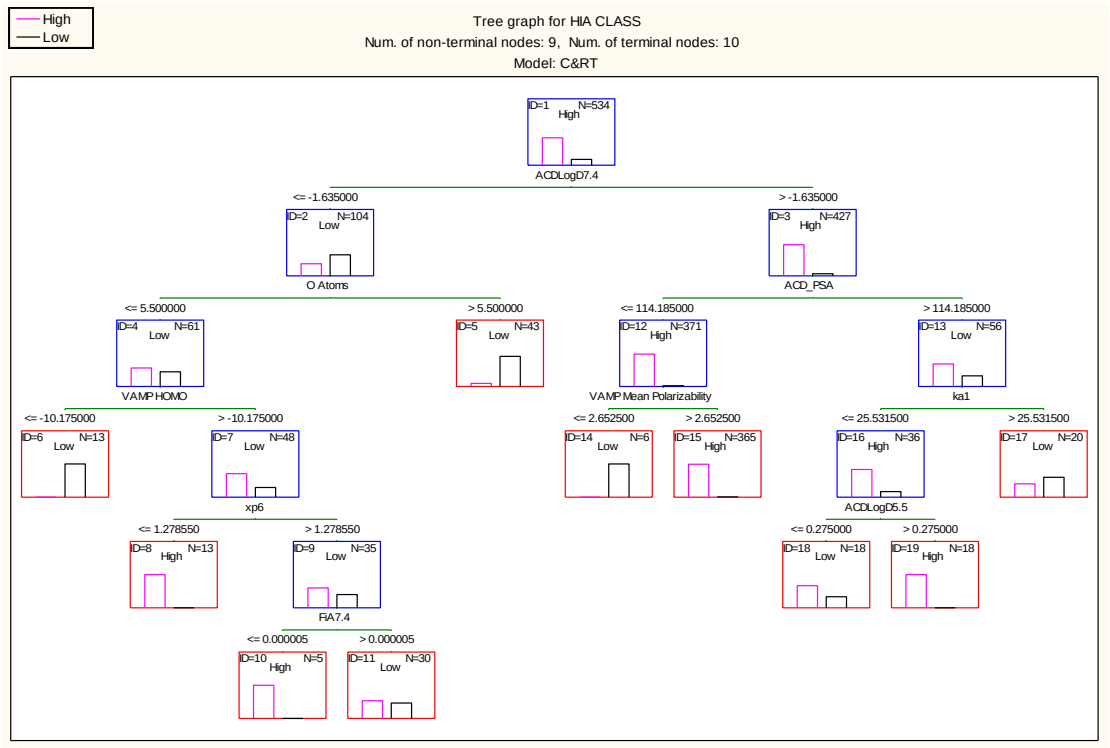

Figure 8.3. Tree graph for C&RT analysis using random forest predictor importance as feature selection method with higher misclassification costs applied to reduce false positives (model 9 in Table 8.4)

8.3.1 Interpretation of the Selected Models (Models 2 and 9)

Both models 2 and 9 have been developed using the 20 most significant molecular descriptors selected by random forest analysis. Although the top 20 molecular descriptors were given as input to the C&RT analysis, not all of the molecular descriptors were used to build the decision trees. The first split variable in both models is ACDLogD7.4; the significance of this descriptor has been described in the

previous chapters (see sections 6.4 and 7.3). For compounds to be split into the high absorption class, LogD7.4 has to be greater than -1.63 according to both models. For compounds with low logD7.4 (≤ -1.63), if they contain more than five oxygen atoms they are classed as poorly-absorbed in this terminal node according to both models. This molecular descriptor is linked to the number of hydrogen bond acceptors, highlighted in Lipinski's rule of five (Lipinski *et al*., 1997) which has been previously described. Examples of poorly-absorbed compounds classed in this node are ceftriaxone and raffinose.

In both models, the next important descriptor selected for the partitioning of compounds with low logD7.4 and less than six oxygen atoms is VAMP HOMO. The higher HOMO energy (> -10.18 in the split in the trees) indicates higher absorption classification as found in the previous chapter 7 (See section 7.3). Compounds with low HOMO values are in the low absorption terminal nodes. The majority of compounds with low HOMO energy (< -10.18) according to this split contain a permanent quaternary ion such as pralidoxime and bethanechol, which are small polar molecules mainly related to the neurotransmitter acetylcholine, or compounds such as fosmidomycin and fosfomycin, which contain phosphorus atoms. Compounds with a higher HOMO energy are further split with different molecular descriptors in the two trees.

In Figure 8.2, these high HOMO compounds are classed as poorly-absorbed if they have more than seven heteroatoms. This corresponds to Lipinski's rule of five, more precisely the number of hydrogen bond acceptors rule. In this node, the majority of compounds are antibiotics such as meropenem and imipenem, which are both poorly-absorbed. There are also some misclassified antibiotics such as penicillin V and amoxicillin, which are highly-absorbed. However, both these compounds have been found to be substrates for the oligopeptide transporter, PEPT1 (SLC15A1), influx transporter in the small intestine (Brandsch *et al*., 2008). The remaining 30 compounds are classed as highly-absorbed if they contain fewer than three OH groups (SsOH_Acnt).

In Figure 8.3 however, such high HOMO compounds are classed as highly-absorbed if they have low xp6 values. The descriptor xp6 is the sixth order single path

molecular connectivity index (Hall and Kier, 1991), which may be regarded as a size descriptor with some shape/connectivity elements. Examples of compounds in this node are those with a small, polar, often peptide like nature with no permanent charge and mainly natural or semi-synthetic compounds such as phenylalanine and captopril, which may have the possibility to be absorbed using oligopeptide transporters (Figure 8.3, Node ID=8). The remaining 35 compounds are classed as poorly-absorbed if they have acidic groups with ionization fraction > 0.000005 at pH7.4.

Highly-absorbed compounds with logD value greater than -1.63 are split differently in Figures 8.2 and 8.3. Despite this, the best molecular descriptors for splitting of these 427 compounds in both trees are the same, namely polarizability (VAMP mean polarizability) and PSA. In both trees, compounds with polarizability values ≤ 2.65 are poorly-absorbed. This molecular descriptor indicates the distortion of a compound's electron cloud by an external electric field (Wang *et al*., 2007). Examples of compounds with ≤ 2.65 polarizability values (Node ID = 12 in Figure 8.2 and 14 in Figure 8.3) are bephenium and vecuronium, both with low polarizability due to the permanent quaternary ammonium ion present in the molecules. Next, PSA is used in both trees, and in both trees compounds with high PSA are poorly-absorbed. In Figure 8.2 a compound is poorly-absorbed if the PSA is greater than 139.67Å, which matches the literature threshold value where it was cited that a molecule will be poorly-absorbed (<10% FA) if the PSA is ≥ 140Å (Palm *et al*., 1997, Clark, 1999). In Figure 8.3, a threshold value of 114.19 Å has been used, but these high PSA compounds (>114.19) have been partitioned further and those with smaller molecular size, as indicated by ka1, and higher logD5.5 values than 0.275 are classed as highly-absorbed. An interesting feature can be observed in Figure 8.2, where for the compounds with PSA values ≤139.67 and low >CH-groups (SsssCH ≤ -1.35), if polarizability is too high (VAMP mean polarizability >56.363) then oral absorption will be poor. Examples of these drugs are two pro-drug ACE inhibitors moexipril diacid and fosinopril plus the cardiac glycoside cymarin.

8.3.2 Chemical Space and Repeating Misclassifications in Models

There were a few compounds that were continually misclassified by most models for the external validation set. The compound lovastatin was misclassified by all models and frovatriptan was misclassified by the majority of models. Both of these compounds are poorly-absorbed, but the models misclassified them as highly-absorbed. Lovastatin is a naturally occurring product used to reduce cholesterol; this compound has poor solubility issues in aqueous medium (Serajuddin *et al*., 1991), plus it has been identified as heavily undergoing gut metabolism, both of which could account for the misclassification (Jacobsen *et al*., 1999). In addition, this compound has been identified as a potential substrate and inhibitor of the efflux transporter P-gp (Wang *et al*., 2001). Frovatriptan, according to the Varma *et al* (2010), has a fraction escaping gut metabolism of 69%; meaning potentially 30% could be metabolised by the gut, specifically UDP-glucuronosyltransferases (UGT's) in the gut due to their substrate specificity of the indole group present in frovatriptan and the similarity of this compound to serotonin, a UGT substrate. However, there is no direct evidence of this in the literature; nevertheless this could explain the misclassification by my models (Varma *et al*., 2010, Krishnaswamy *et al*., 2003).

8.4 Discussion

In this chapter I used various filter feature selection methods for data pre-processing, to pick significant descriptors related to intestinal absorption. These descriptor sets were used as input for C&RT analysis, which has an embedded feature selection method, to classify compounds into high or low absorption in an unbalanced-class dataset – sometimes called a "biased" dataset. The application of higher misclassification costs for false positives to the C&RT analysis was also investigated to overcome the problem of biased datasets (which contain many more highly-absorbed compounds than poorly-absorbed compounds) and to see if models with greater predictive accuracy could be achieved.

The feature selection methods used in this chapter were predictor importance using random forest (RF), chi square (CS), information gain ratio (IGR), greedy search (GRD) and genetic search (GEN). The feature selection methods were compared based on the predictive ability of the C&RT algorithm. There were certain

expectations of the feature selection methods based on how they work and their advantages and disadvantages. To begin, it was expected that the combination of a pre-processing feature selection method and C&RT, which has an embedded feature selection, would have higher predictive accuracy when compared to using C&RT with no pre-processing feature selection method. This was on the basis that when C&RT splits compounds, further down the tree there are fewer compounds in the deeper nodes, therefore less statistical support for an effective selection of the best descriptor, especially when there are a larger number of molecular descriptors to choose from. Therefore, as a result the C&RT algorithm could pick descriptors that may be less relevant to molecular descriptors higher up in the tree. However, C&RT is a successful technique in its own right with an embedded feature selection function which is used in model development for the prediction of oral absorption (see chapter 7) (Newby *et al*., 2013a, Deconinck *et al*., 2005). The benefits of using C&RT are that it can cope with noisy data (to some extent) of moderately sized biased datasets and produces models (decision trees) that in principle can be easily interpreted (Suenderhauf *et al*., 2011). In addition, it is less time consuming than pre-processing the molecular descriptors first.

The expectations of the feature selection methods themselves can be considered and compared to the obtained results in this chapter. The benefits of simple univariate filter techniques such as CS and IGR are that they are simple and fast to compute; however they fail to take into account molecular descriptor interactions (Saeys *et al*., 2007, Guyon and Elisseeff, 2003). This is in contrast to GRD and GEN, which take molecular descriptor interactions into account but are more computationally expensive. In a comparison of GRD and GEN, due to the way these feature selection methods work, GEN should achieve higher accuracy, as it performs a global search in the molecular descriptor space, whilst GRD performs a local search in the molecular descriptor space. Using the predictor importance in the random forest method is computationally expensive. However, there is the added advantage that misclassification costs can be applied using the software as well as being applied for the C&RT analysis. Finally, based on the previous chapter 7, the application of higher misclassification costs to false positives will produce models with increased

overall accuracy and reduced false positive misclassifications, therefore overcoming the problem of biased datasets compared with equal misclassification costs.

Overall, one of the best feature selection methods according to the models produced in this work was predictor importance using random forest. This was expected for this method, as it was possible to apply higher misclassification costs to the feature selection technique itself as well as applied to the C&RT analysis. Even when misclassification costs were not applied to predictor importance, the produced models still had higher overall accuracies over most models. This is down to the ensemble nature of this method, which is known to perform better than single tree analysis (Dietterich, 2000). In comparison with C&RT, where no pre-processing feature selection was utilised, the random forest predictor importance feature selection method had higher overall accuracy for the validation set in all cases. The high classification accuracy on the training set but low predictive accuracy on the validation set could indicate overfitting of the models produced by C&RT even after pruning. Models produced by other pre-processing feature selection techniques were better compared with models produced by C&RT with no pre-processing feature selection on the validation set, except for the models produced by CS feature selection. In the majority of the cases, using C&RT alone gave better predictive accuracy for the parameter optimisation set compared with IGR, GRD and GEN; however, these latter methods had better overall predictive accuracy for the validation set. This shows that C&RT without pre-processing can cope with redundant and meaningless molecular descriptors; however it is prone to overfitting (even with pruning of the trees) and can lack predictive accuracy in the validation set.

Comparing the expectations set out initially, it was found that, comparing univariate methods such as CS and IGR with those that take into account molecular descriptor interactions (GEN and GRD), there is no clear pattern in the difference between their results. However, overall, when equal misclassification costs were applied to the C&RT analysis, GEN as expected had better or comparable predictor performance relative to CS and IGR. On the other hand, GRD had comparable performance with CS and weaker compared with IGR. When misclassification costs were applied to C&RT, GEN and GRD models were better than CS and IGR for the training and

validation sets; again it is difficult to state which method is better overall. This effect is also seen in the next example when comparing GEN and GRD feature selection methods based on the predictive accuracy of the C&RT analysis. GEN performs better than GRD when equal misclassification costs were used; this matches the predictions previously made. This is in some agreement with work using numerical regression analysis (Xu and Zhang, 2001). However, upon applying higher misclassification costs the molecular descriptors pre-processed by the GRD model outperformed the GEN model. This could be due to the correlation-based feature selection subset evaluator used by the GEN method not being suitable for use with C&RT and misclassification costs, and potentially highlight overfitting by the GEN based model. In this chapter the application of higher misclassification costs to false positives resulted in better overall accuracy and specificity as expected in the majority of cases and confirms the results from the previous chapter 7.

In this chapter I have shown that for most models using pre-processing feature selection does appear to improve classification accuracy compared to the control (C&RT using all molecular descriptors) based on predictive accuracy. This agrees with work carried out by Xue and co-workers (Xue *et al*., 2004), who considered three different datasets including prediction of oral absorption. They used recursive feature elimination (a type of backwards feature elimination) for feature selection combined with SVM to classify compounds. They compared the results with and without the feature selection method and found that, for oral absorption, improved accuracy was obtained when the feature selection method was used. For one of the datasets, feature selection gave comparable predictive ability, which with a smaller descriptor subset will increase the interpretability of resulting models. It must be noted that unlike C&RT employed in this work, SVM does not have an embedded feature selection capability. A related work is a study by Suenderhauf *et al* (Suenderhauf *et al*., 2011) who carried out regression and classification for oral absorption using a variety of techniques including C&RT, Support Vector Machine (SVM), and chi-squared automatic interactor detector (CHAID). Using these classification techniques, they compared the feature selection methods of best first feature selection (BFS) using a greedy hill-climbing algorithm, linear correlation analysis and decision tree splitting criteria. Suenderhauf utilised the decision trees to

pick smaller subsets of molecular descriptors used in the work as input for the model development. This is similar idea to the proposed in this thesis of a two-stage pre-processing feature selection. The best model was produced by CHAID using the entire set of original molecular descriptors, which contradicts my results that pre-processing feature selection gives better accuracy. However, it is interesting to note that, out of the feature selection methods that they used, the decision tree splitting criteria gave the best results. This research also showed that SVM had poor performance when utilised with feature selection methods used in the study (unlike the study by Xue *et al*. 2004), therefore this could indicate that some feature selection methods work better than others with SVM. This has been supported by a recent study comparing different learners and feature selection methods (Eklund *et al*., 2014). What is apparent with other feature selected QSAR studies in the literature is that different feature techniques need to be tested and tried. In particular in a recent study, it was shown that modelling methods that cannot handle larger numbers of molecular descriptors such as MLR will benefit from feature selection to reduce the number of molecular descriptors (Eklund *et al*., 2014). In this chapter I have shown that even though C&RT can handle a large number of features, overall, it benefits from using pre-processing feature selection to improve model accuracy.

Although it is difficult to directly compare the different feature selection techniques that I used with the literature, the molecular descriptor subsets can be compared. Firstly it is interesting to compare in this work the molecular descriptors selected by the pre-processing feature selection methods (Appendix 2, Tables A2.1 and A2.2). The top molecular descriptors picked by the feature selection methods can be found in Table 8.5. This table shows the top molecular descriptors that were picked by three or more feature selection methods. The molecular descriptors selected by the various feature selection methods were used as input for C&RT analysis, which in turn further selected a smaller subset of molecular descriptors to build decision trees.

Table 8.5. Molecular Descriptors Selected By Three or More Pre-Processing Feature Selection Methods Listed in Table 8.2

Descriptor	Feature selection method	Description
ACDLogD7.4	RF, RF (MC), CS, IGR, GRD, GEN	Apparent distribution coefficient at pH 7.4 calculated by ACD
ACDLogD10	RF, CS, IGR, GRD, GEN	Apparent distribution coefficient at pH 10 calculated by ACD
ACDLogD5.5	RF, RF (MC), CS, GRD, GEN	Apparent distribution coefficient at pH 5.5 calculated by ACD
SHHBd	CS, IGR, GRD, GEN	Sum of the hydrogen atom level E-state values for all hydrogen atoms bonded to donating atoms
O Atoms	RF, RF (MC), CS, GRD	Number of oxygen atoms in whole molecule
ACD_PSA	RF, RF (MC), CS, GRD	Polar surface area
numHBa	RF, RF (MC), CS, GEN	Number of Hydrogen bond acceptors
SsOH_acnt	RF, RF (MC), CS, GEN	Counts of atom-type E-state for hydroxyl groups
VAMP Heat of Formation	RF, RF (MC), GRD, GEN	Enthalpy required to form 1 mole of compound at 298K calculated by VAMP
ACD_LogP	CS, GRD, GEN	Octanol/water partition coefficient calculated by ACD
ACDLogD6.5	RF, CS, GRD	Apparent distribution coefficient at pH 6.5 calculated by ACD
Heteroatoms	RF (MC), CS, GEN	Number of atoms that are not carbon or hydrogen e.g. nitrogen, oxygen
ka1	RF, RF (MC), GEN	First order kappa alpha shape index
numHBd	RF, CS, GEN	Number of hydrogen bond donors
SdsssP	IGR, GRD, GEN	Sum of atom-type E-state for phosphorous atoms with 3 single and one double bond
Sum of E-State indices	RF, IGR, GEN	Sum of the E-State values for all the atoms in molecule
VAMP HOMO	RF (MC), GRD, GEN	Energy of the highest occupied molecular orbital calculated by VAMP
VAMP LUMO	RF, GRD, GEN	Energy of the lowest occupied molecular orbital calculated by VAMP

The top descriptors picked by firstly the pre-processing method and then by C&RT analysis are shown in Table 8.6. Table 8.6 also indicates the number of times a molecular descriptor was picked by C&RT with or without pre-processing feature selection.

Table 8.6. The Top Molecular Descriptors Selected by C&RT

		Number of C&RT models	
Type of descriptor	**Descriptor**	**Pre-processing**	**No Pre-processing**
Hydrogen bonding	ACD_PSA	8	1
	O Atoms	8	1
	SHHBd	8[a]	
Lipophilicity	ACDLogD7.4	10	1
	ACD_LogP	6	
	ACDLogD6.5	3	1
Polarity/ Polarization	VAMP LUMO	4[a]	2
	N+	5[a]	
	VAMP Mean Polarizability	5[a]	
Size/Shape	VAMP totl Energy	5[a]	
	ka1	3[a]	
	SsssCH	3	1

[a]Occurred more than once in a single tree model.

The top molecular descriptor picked by the majority of feature selection methods was the same as the top molecular descriptor then picked by the resulting C&RT analysis (ACDLogD7.4). Other studies have identified lipophilicity descriptors, in particular logD7.4 as well as logD5.5, logD6.5 and logP, as important for intestinal absorption, as picked by various feature selection techniques (Suenderhauf *et al*., 2011, Agatonovic-Kustrin *et al*., 2001, Winiwarter *et al*., 1998). The next most frequently picked molecular descriptors are those relating to hydrogen bonding, in particular PSA. The importance of PSA has been emphasised by its constant selection throughout the previous chapters of this thesis. This descriptor is used in many literature models for oral absorption as well as those studies which focus on feature selection methods for oral absorption (Wegner *et al*., 2004). The other top hydrogen bonding descriptors highly ranked are the number of oxygen atoms and SHHBd, which is related to the number of hydrogen bond donors in a molecule. Both these descriptors were picked by the feature selection models and utilised in the C&RT analysis high up near the tree root, indicating the importance of these descriptors. Descriptors relating to hydrogen bonding capacity are important in oral absorption modelling and are used in the widely accepted filter, Lipinski's rule of five as described in previous chapters (Lipinski *et al*., 1997). Overall the top descriptors picked by the feature selection methods and then utilised by C&RT are very similar. Also, the majority of molecular descriptors used by C&RT without any pre-processing feature selection match those picked by the pre-processing feature selection methods, with a few exceptions. The top descriptors in Table 8.6 are in line

with the literature, where among these molecular descriptors related to absorption are those that describe lipophilicity, molecular size/shape, polar surface area, hydrogen bonding, and similar parameters.

8.5 Conclusion

Feature selection is important in its many forms as a way to increase interpretability and predictive accuracy, as well as reducing over-fitting of QSAR models. This chapter has shown that pre-processing filter feature selection methods can greatly improve QSAR models using C&RT analysis. C&RT can be used as an embedded feature selection method; however, it can be inadequate since further down the tree, there are fewer compounds available for descriptor selection and therefore descriptors may be selected which are not optimal. Here, I have used several pre-processing feature selection methods prior to C&RT and have produced more accurate QSAR models for the estimation of oral absorption class, as shown by the external sets of compounds. However, examination of the literature reveals that different feature selection methods utilised with different classification methods should be tried and evaluated. Similar molecular descriptors were picked by the different feature selection methods; and those descriptors relate to lipophilicity, hydrogen bonding, polarity, size and shape. Higher misclassification costs applied to reduce false positives yielded models with better overall predictive accuracy of highly and poorly-absorbed compounds. The use of filter pre-processing feature selection methods and misclassification costs produce models with better interpretability and predictive accuracy that overcome the problem of a biased (unbalanced-class) dataset with many more highly-absorbed compounds than poorly-absorbed compounds, and shows the importance of feature selection in QSAR model development.

9 Decision Trees to Characterise the Roles of Permeability and Solubility on the Prediction of Oral Absorption

9.1 Introduction

The fundamental properties that govern intestinal absorption are permeability and solubility (Amidon *et al*., 1995). It is well understood that a compound must be soluble in intestinal fluid in order to be absorbed. Therefore, solubility can affect the oral absorption of compounds. There are an increasing number of poorly soluble compounds that are introduced to the market as new drug candidates; therefore the prediction of New Chemical Entities (NCEs) is problematic. In spite of this, there are few studies that incorporate both experimental solubility and permeability values within one single model, in order to see the effect these two properties have on oral absorption (Pade and Stavchansky, 1998, Bergstrom *et al*., 2003). Instead, most studies have removed compounds with solubility issues when modelling oral absorption (Zhao *et al*., 2002, Hou *et al*., 2007c), which is not ideal due to the increasing number of poorly soluble drugs being developed.

Even if models were to include experimental solubility, another issue here is the lack of experimental solubility for drug compounds to be used in oral absorption modelling. Solubility itself is a complex parameter and in turn dependent on numerous factors. Therefore, it is important to investigate what multiple elements, such as those calculated from the molecular structure, may improve understanding of this property in relation to absorption. Additionally, it is important to determine suitable alternatives for solubility which could act as a surrogate if experimental solubility was not available, and their impact on absorption prediction. For example there are solubility models calculated via GSE (Jain and Yalkowsky, 2001), melting point, dose number and MPbAP (Chu and Yalkowsky, 2009).

Based on the literature, a large dataset is needed in order to see the effects of solubility and permeability on the fraction absorbed. Therefore, the first aim of this chapter was to expand the permeability dataset by combining data from Caco-2 and MDCK cell lines. By studying the linear relationship and the effect of different absorption mechanisms between the two cell lines and from the differences already

known between the two cell lines, the justification of combining the datasets can be shown. The justification can be confirmed based on the literature utilising (smaller) permeability datasets (Irvine *et al*., 1999). Secondly, the determination of a permeability threshold to predict fraction absorbed class using an objective decision tree method is tested on an external validation set. This is justified as the majority of literature models determine this threshold subjectively (Artursson and Karlsson, 1991). Using this permeability threshold, experimental and predicted solubility and related properties such as dose number and melting point were incorporated, in addition to structural molecular descriptors, to build decision tree-based classification models to predict HIA class. Based on this chapter, one can obtain an increased understanding around the relationship between two popular cell based assays and how they can be used to predict absorption class using an objective permeability threshold. In addition, the effect of solubility and related properties on the models for the prediction of fraction absorbed is explored.

9.2 Methods

9.2.1 Dataset

Dataset 3 (as previously described in the Datasets and Methods section 5.1.3) was used for this chapter and contained collected data for %HIA, *in vitro* apparent permeability, aqueous solubility, maximum dose strength and melting point. In addition to these collected values from dataset 3, predicted solubility, dose number and Melting Point based Absorption Potential (MPbAP) were calculated using the equations below.

9.2.1.1 Predicted Solubility

Solubility was calculated by the revised general solubility equation (GSE) using collected experimental melting point and calculated logP (Jain and Yalkowsky, 2001) (Equation 9.1).

$$\text{Log Solubility (GSE)} = 0.5 - 0.01\,(MP - 25) - \text{LogP} \qquad \textbf{Eq. 9.1}$$

9.2.1.2Dose number

Dose number is a dimensionless number used to determine high or low solubility in the Biopharmaceutical Classification System (BCS) (Amidon *et al*., 1995). It is calculated using the solubility and maximum strength dose (Equation 9.2).

$$D_o = (M_o / V_o) / S \qquad \textbf{Eq. 9.2}$$

Where D_o is dose number, M_o is the highest dose strength, V_o is 250ml and S is the aqueous solubility (mg/ml). The maximum strength dose was obtained from the literature as explained in the Datasets and Methods section 5.1.3.

9.2.1.3Melting Point Based Absorption Potential

The melting point based absorption potential (MPbAP) was derived from the GSE but utilising maximum dose as well as melting point (Chu and Yalkowsky, 2009), as shown by Equation 9.3.

$$\text{MPbAP} = 0.5 - 0.01\,(\text{MP}-25) - \log(4 * \text{Maximum Dose}) \qquad \textbf{Eq. 9.3}$$

9.2.2 Training Sets and Validation Sets

Using the combined permeability data from the two cell lines yielded an initial dataset of 447 compounds. Compounds with MDCK and Caco-2 permeability data that differed by more than one log unit and one compound that did not have a numerical value for HIA were removed (14 compounds in total). This resulted in a dataset of 433 compounds. The 433 compounds were split into a training set and a validation set. To ensure a similar distribution of fraction absorbed in these two sets, compounds were sorted according to ascending %HIA and then logP values. From each group of six consecutive compounds, five were assigned to the training set, and one compound was allocated to the validation set randomly. The initial training set consisted of 356 compounds and the validation set consisted of 77 compounds.

For models used to determine the influence of solubility and related parameters, compounds that had missing values for solubility, melting point and dose number were removed from the initial training and validation sets. The final compound numbers for decision tree analysis in this chapter are shown in Table 9.1.

Table 9.1. Compound numbers used in the training and validation sets for decision tree analysis

Property	Total number of compounds	Training set n	Validation set n
Permeability	433	356	77
Solubility	296	242	54
GSE solubility	315	262	53
Dose number	292	239	53
Melting point	315	262	53
MPbAP	308	257	51

9.2.3 Model Development

9.2.3.1 Molecular Descriptors

Molecular descriptors were calculated from structures using the software packages TSAR 3D v3.3 (Accelrys Inc.), MDL QSAR (Accelrys Inc.), MOE v2010.10 (Chemical Computing Group Inc.) and Advanced Chemistry Development ACD Laboratories/LogD Suite v12. Including the seven descriptors of permeability, solubility and related parameters, a total of 220 molecular descriptors were utilised for analysis.

9.2.3.2 Permeability Threshold Determination Using C&RT

The permeability threshold is the numerical value chosen by C&RT that best predicts HIA class. In this chapter several different analyses were performed where high absorption compounds were defined as those having HIA values of above 30, 50, 70, 80 or 90%. Using the training set of 356 compounds, HIA class was used as the dependent variable and permeability as the independent variable. The C&RT analysis was restricted to only one split to give the permeability threshold. This threshold was tested using a validation set of 77 compounds. Due to the class imbalance, where there are many more highly-absorbed than poorly-absorbed compounds, higher misclassification costs were applied to false positives to overcome this class distribution bias. Based on chapters 7 and 8, the use of misclassification costs has shown improved model accuracy. The misclassification cost values applied depended on the class distribution of the dataset. For instance,

when the "high absorption" class is defined as having %HIA ≥ 30%, the cost of a false positive was considered five times the cost of a false negative due to roughly five times more highly-absorbed compounds in the dataset. Misclassification costs of 5, 4, 3, 2.5 and 2 were applied to false positives in the analyses where the high HIA class had been defined as those compounds having %HIA values equal or above 30, 50, 70, 80 and 90%, respectively.

9.2.3.3Permeabilityand Solubility Related Model Analysis for Oral Absorption Class Determination

In this section, models were built using HIA class as the dependent variable where high absorption was defined as HIA ≥ 80% and molecular descriptors were utilised as the independent variables for model building. The HIA class definition of ≥ 80% was selected based on preliminary work, where when using a lower HIA class definition such as 30-70% only poor models could be achieved, due to the lower number of poorly-absorbed compounds. Using a higher threshold of 90% resulted in poorer overall accuracy (based on preliminary analysis), and this threshold is too high to predict oral absorption class effectively with a high number of false negatives.

In this chapter, permeability was set as the first split variable and two alternative approaches were used to choose the remaining split variables. In the first one, the C&RT tree was allowed to grow automatically. In the second one, each of the solubility and related parameters (dose number and melting point) were manually chosen as the second split variables (note that C&RT still chooses the cut-off point automatically) and then the tree was allowed to grow automatically. Stopping factor, in particular the minimum number of compounds for splitting, were used to prevent overfitting of the C&RT trees. This minimum number was set at 11 for the permeability only C&RT trees (with additional molecular descriptors but no solubility related parameter) and eight for permeability and solubility trees (again with additional molecular descriptors).

To determine the relationship between Caco-2 and MDCK permeability, MINITAB Statistical Software (version 16.1.1.0) and Prism (GraphPad Software, Inc) v.5.02 were used to carry out linear regression, identify outliers and perform statistical significance testing between the different absorption mechanisms. For linear regression the parameter reported to assess the fit of the two variables (permeability in Caco-2 vs. permeability in MDCK) was the squared correlation coefficient, r^2 forced through the origin. For correlation analysis the Pearson's correlation coefficient (r_p) and the Spearman's ranking correlation coefficient (r_s) were calculated. It must be emphasised here that r^2 based on the regression line forced through the origin is not comparable to r^2 values where the regression line is not forced through the origin (Hahn, 1977). The statistical significance of the correlations and regression lines and comparison of the regression lines for different absorption mechanisms (using the intercept and the slope values) was depicted by p values. P values < 0.05 indicated significance.

The predictive performance of the classification models built using C&RT was measured using the same measures used in chapter 7 and 8 apart from CNMI due to the different misclassification costs applied to different parts of the C&RT analysis.

9.3 Results and Discussion

In this chapter, in order to investigate the effects of permeability and solubility, a large dataset of human intestinal absorption was gathered from the original literature and then for the same compounds, Caco-2 and MDCK permabilities, solubility, melting point and dose were gathered from the original literature. This dataset was collected in order to develop models for predicting high/low oral absorption and to explore the suitability of different solubility and permeability measures from different sources as descriptors of intestinal absorption.

In terms of permeability, I have gathered permeability measured in both Caco-2 and MDCK cell lines. *In vitro* permeability through different cell lines is commonly used as a high throughput measure of effective intestinal absorption in early drug discovery. There have been a few studies which show the linear relationship between these cell lines. For example, Braun *et al* (Braun *et al*., 2000) studied the relationship

between Caco-2 and MDCK cell lines and from 14 compounds achieved an r^2 of 0.86. However, Avdeef and Tam (Avdeef and Tam, 2010) achieved a r^2 of 0.90 using a dataset of 79 compounds.

9.3.1 Comparison of Caco-2 and MDCK Apparent Permeability as Indicators of Intestinal Absorption

For 185 compounds, the *in vitro* apparent permeability from both Caco-2 and MDCK cell lines was obtained from the literature. By an exhaustive literature search possible transport routes were identified for all these compounds. Plotting the permeability of these two cell lines on a log scale a linear relationship is shown (Figure 9.1) where the transport routes have also been highlighted. Out of 185 compounds in this figure, 96 compounds were found to be substrates of at least one transporter system and 11 compounds have been suggested to be absorbed to some extent via paracellular route.

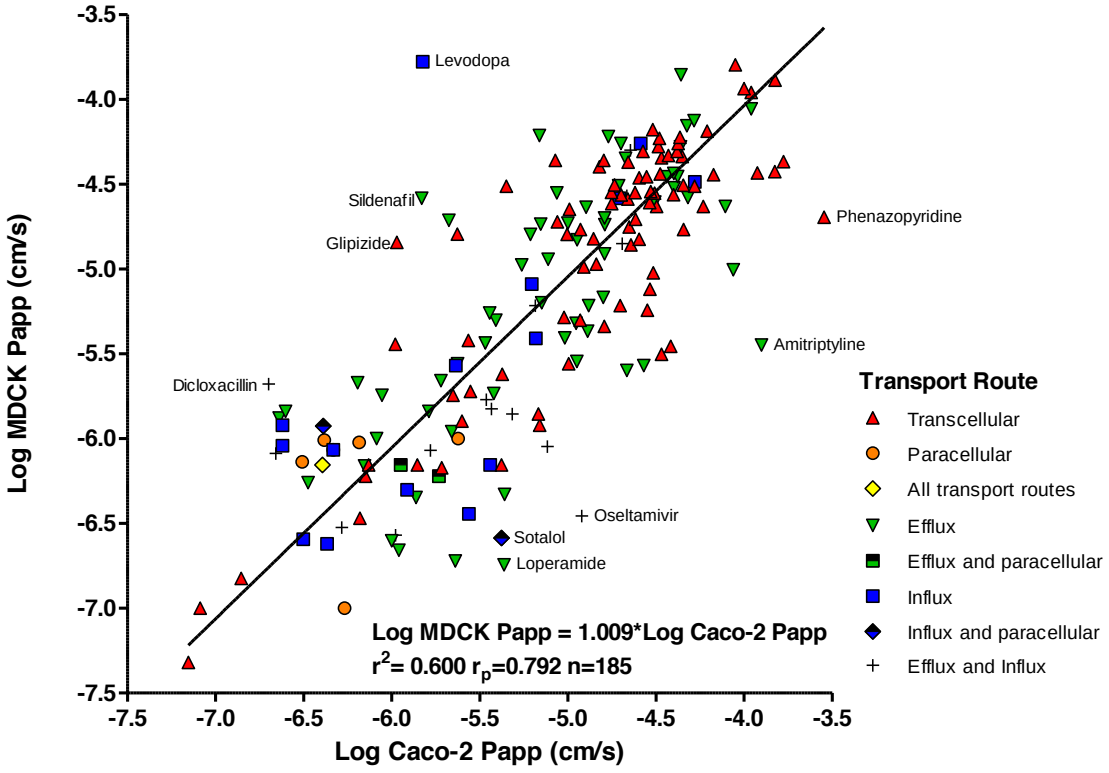

Figure 9.1. Linear relationship between Caco-2 and MDCK apparent permeability for 185 compounds

It can be seen in the plot that Caco-2 and MDCK permeability of the majority of compounds correlate well with each other regardless of their absorption routes. However, there are compounds that deviate significantly from this line, and the removal of 9 outlier compounds (compound names shown in the figure) improves the correlation significantly (Table 9.2). Details of the outlier compounds and a description of reasons for removal can be found in Appendix 3 (Table A3.1). A better linear relationship between the two cell lines is also achieved when only compounds undergoing passive transcellular absorption are plotted (Table 9.2). It can be noted in Table 9.2 that the correlations between the cell lines are better after the removal of 9 outliers than after the removal of all the compounds with a transporter effect. It is also noteworthy that not all the outliers were substrates of a transporter; examples are phenazopyridine and glipizide, where no transport system other than passive-transcellular has been identified. Both these drugs have poor solubilities (dissolution limiting solubility) and are classed in Class II of Biopharmaceutics Classification System (BCS) (Gao, 2012, Mehramizi *et al*., 2007).

Similar conclusions can be made from the results of previous studies where transporter mediated effects could not be identified by correlating the permeability through different cell lines. Irvine *et al* (Irvine *et al*., 1999) compared the apparent permeability of 55 compounds using MDCK and Caco-2 cells. This study achieved an r^2 of 0.79. Irvine identified 12 compounds that were substrates for carrier mediated systems. I crossed referenced the remaining compounds used by Irvine with my database and identified an additional set of 18 compounds to be substrates for carrier mediated systems. Therefore, over half of this original dataset has now been found to be affected by a carrier mediated route. The 12 compounds highlighted as undergoing carrier systems in most cases were within the linear fit of Irvine's, with only a few exceptions. The explanation by Irvine of why known P-gp substrates were not identified when comparing the two cell lines is not suitable. For the P-gp substrates highlighted in Irvine's work, it was stated the reason they could not be identified was due to saturation of the transport mechanism in the assay. However, Braun *et al* (Braun *et al*., 2000) used the same compounds but at lower concentrations, and they were still unable to identify known P-gp substrates. It was concluded that using the relationship between MDCK and Caco-2 could not identify

P-gp substrates. From this work the correlation between MDCK and Caco-2 permeability does indicate the same result that compounds with carrier mediated mechanisms do not deviate from the correlation between Caco-2 and MDCK permeabilities. This is despite the fact that the transporters have different abundance levels in these two cell lines.

Table 9.2. Statistical parameters for the linear relationship between MDCK and Caco-2 permeability measured using PRISM

Datasets	**r^2 (with intercept)**	**r^2 (no intercept)**	**R_p**	**R_s**
All compounds (185)	0.63	0.60	0.79	0.79
Passive transcellular (83)	0.71	0.67	0.84	0.74
OUTLIERS Removed (9 removed)				
All compounds (176)	0.73	0.72	0.86	0.84
Passive transcellular (81)	0.75	0.75	0.87	0.76

A table was compiled that compares the cells and small intestine in terms of species origin, tightness of the cell junctions and also the transporter and enzyme expressions (Appendix 3, Table A3.2). One thing to note is the lack of information or evidence in the literature for transporter and enzyme expression, especially for the specific strains of the MDCK cell line which is less well studied. For the small intestine the expression of transporters and enzyme systems can vary from the three sections of the small intestine, as compounds are not just absorbed from one section, hence I tried to accommodate an overview of expression from the human small intestine (Englund *et al*., 2006). It can be seen from Table A3.2 in Appendix 3 that the main differences between MDCK and Caco-2 cell lines in general are that MDCK does not express some transporter types and that MDCK has a lower abundance of some of the other transporters compared to Caco-2 cell lines. However it must be noted that expression of transporters or enzymes does not necessarily correlate with their functionality for affecting the absorption of the compounds across different membrane/cell lines (Ungell, 2004, Hilgendorf *et al*., 2007), and as it was shown earlier, most substrates of different transporters do not deviate from the correlation between Caco-2 and MDCK permeabilities.

The different expression levels of metabolising enzymes in the different cell lines could also potentially affect the permeability of compounds. The expression and

activity of CYP3A4 enzymes in Caco-2 cells are either not present or very weak (Le Ferrec *et al*., 2001, Hayeshi *et al*., 2008). A recent investigation has found no evidence of CYP3A4 expression in MDCK II cells (Quan *et al*., 2012). Unfortunately the lack of information regarding enzymatic activity in the cell lines makes it difficult to comprehensively compare and contrast the suitability of these *in vitro* tools as indicators of intestinal absorption.

Cell based assays, particularly Caco-2, have a reputation for variability. The differences can arise from the experimental conditions, which in turn can affect the monolayer, those that affect the analysis of samples and also the physioc-chemical properties of the compound (Shah *et al*., 2006). A good example is solubility, which depending on experimental conditions can cause variation particularly for compounds with low solubility such as the outlier compounds phenazopyridine and glipizide (Mehramizi *et al*., 2007, Gao, 2012) (Figure 9.1).

The prime purpose of cell based assays such as Caco-2 and MDCK is to study the rate of passive permeability rather than other transport routes involving influx and efflux transporters. In this dataset, out of the 185 compounds, 96 were identified as undergoing transport routes other than passive. In some cases, more than one route was identified as being involved for the transport of the compound (Table 9.3).

Table 9.3. The different identified absorption mechanism of the 185 compounds

Transport route	**Number of compounds**	**Examples**
Passive transcellular (A)	83	sumatriptan, valsartan
Passive paracellular (B)	6	lucifer yellow, mannitol
Efflux (C)	62	vinblastine, saquinavir
Efflux and paracellular (D)	2	famotidine, cimetidine
Influx (E)	15	amoxicillin, tolbutamide
Influx and paracellular (F)	2	soltalol, atenolol
Efflux and influx (G)	14	talinolol, acebutolol
Influx, efflux & paracellular (H)	1	Ranitidine

From Table 9.3, there are a higher number of compounds identified as carrier mediated efflux substrates compared to influx substrates. The majority of compounds that were identified as efflux substrates are substrates of the P-gp transporter, which is always tested due to the great influence this transporter has on reducing absorption of many compounds.

I compared the permeability values obtained from Caco-2 and MDCK cell lines for all compounds and subgroups of compounds showing specific routes of absorption as described in Table 9.5. Two statistical methods were employed; 1) paired student t-test to compare MDCK and Caco-2 permeability values of a subgroup of compounds, and 2) comparison of the coefficients of the correlation lines of subgroups of compounds, e.g. efflux substrates and compounds with passive transcellular absorption. The results for subgroups indicated that permeabilities through MDCK and Caco-2 cell lines are correlated with similar slopes and intercepts for compounds with different absorption mechanisms (Table A3.3 and Figures A3.1-A3.7 in the Appendix 3). The only significant difference between the correlation lines was the difference between compounds undergoing transcellular and paracellular absorption routes (p value 0.0023). However, despite the different tightness of the Caco-2 and MDCK cell lines, the observed difference may be due to the narrow range of permeability values of the compounds with paracellular absorption route resulting in a non-significant correlation between MDCK and Caco-2 solubility of this subgroup (Figure A3.1 in Appendix 3). This hypothesis is supported by the results of a paired student t test between the permeability values of the two cell lines for the 11 compounds undergoing paracellular absorption (as a main or shared transport route), which showed no significant difference between Caco-2 and MDCK permeabilities (p value > 0.05). In addition, paired t tests for all different absorption mechanism groups were made and no significant differences between the two cell lines for these absorption groups were found. Therefore, it can conclude that in general there are no statistically significant differences between the two cell lines even when considering separately the compounds with different absorption mechanisms. Therefore, the data from both these cell lines can be combined into a larger permeability dataset for use in further modelling.

9.3.2 Determining Permeability Threshold for an Effective Oral Absorption

In this work I use the large dataset of combined Caco-2 and MDCK permeability and a statistical method (C&RT) to identify statistically valid permeability threshold for high/low oral absorption. Using C&RT analysis, a permeability threshold value was obtained to predict the high or low intestinal absorption (HIA class) using a training set of 356 compounds. Several different analyses were performed where high

absorption compounds were defined as those having HIA values of above 30, 50, 70, 80 or 90%. In order to optimise the threshold selection, various C&RT models using different misclassification cost ratios for false positives: false negatives (FP:FN) were generated. The results show the permeability threshold selected by the C&RT analyses and the accuracy, specificity and sensitivity of the class prediction (Table 9.4).

Table 9.4. The permeability thresholds selected by C&RT and HIA class prediction with equal and higher misclassification costs applied to false positives when high HIA is defined as higher than 30, 50, 70, 80 and 90%

Model	HIA class determination above or below	Set	Misclassification Costs (FP:FN)	SP X SE	SE	SP	Log Perm Threshold	Perm Threshold (cm/s x10^{-6})
1	30%	t	1:1	0.000	1.000	0.000	-6.11	0.78
		v		0.000	0.986	0.000		
2	50%	t	1:1	0.626	0.905	0.692	-6.02	0.96
		v		0.470	0.939	0.500		
3	70%	t	1:1	0.562	0.910	0.618	-5.91	1.23
		v		0.522	0.948	0.550		
4	80%	t	1:1	0.645	0.745	0.865	-5.15	7.08
		v		0.630	0.741	0.850		
5	90%	t	1:1	0.565	0.785	0.720	-5.08	8.32
		v		0.487	0.762	0.639		
6	30%	t	5:1	0.672	0.874	0.769	-5.98	1.05
		v		**0.800**	**0.914**	**0.875**		
7	50%	t	4:1	0.664	0.803	0.827	-5.64	2.29
		v		0.720	0.864	0.833		
8	70%	t	3:1	0.645	0.745	0.865	-5.15	7.08
		v		0.630	0.741	0.850		
9	80%	t	2.5:1	0.645	0.745	0.865	-5.15	7.08
		v		0.630	0.741	0.850		
10	90%	t	2:1	0.566	0.759	0.745	-5.00	10.0
		v		0.533	0.738	0.722		

t-training; v-validation; Sensitivity is equivalent to the number of correctly classified highly-absorbed compounds and is calculated using SE=(TP/(TP+FN)); Specificity is equivalent to the number of correctly classified poorly-absorbed compounds and is calculated using SP=(TN/(TN+FP)); TP-true positive; FN-False negative; TN-true negative; FP-false positive; Overall Accuracy of the models was calculated by multiplying Specificity by Sensitivity (SP x SE)

It can be seen in Table 9.4 that using high ratios of (FP:FN) misclassification costs resulted in significantly improved accuracy of the permeability threshold for classification of compounds into high or low absorption groups for all definitions of HIA class. For example using equal misclassification costs to find permeability threshold for dividing compounds into ≥ 30% or < 30% HIA was not successful at all (Model 1 Table 9.4). On the other hand, increasing the cost of false positives to five times that of the false negatives resulted in a high accuracy of classification and a robust threshold of -5.98 (in log units) (model 6) for this classification. It must be noted here that different high/low definitions of HIA resulted in different proportions of compounds in "high" or "low" absorption classes, and hence the choice of misclassification cost ratios to reflect the ratios of highly-absorbed to poorly-absorbed compounds as defined in chapters 7 and 8. Therefore by applying higher misclassification costs to reduce false positives, this shifted the permeability threshold in order to reduce the number of false positives (Figure 9.2). The higher misclassification for false positives is justified due to the under representation of the poorly-absorbed class. The one exception to this is the 80% HIA class definition, where applying misclassification costs had no effect on the permeability threshold. In practice, when using the permeability threshold to classify high/low absorption compounds, the suitable threshold suggested by models 6-10 can be used for HIA class definition. The permeability thresholds determined by C&RT when applying higher misclassification costs from Table 9.4 are shown below in Figure 9.2, when plotting fraction absorbed against permeability for the training and validation sets.

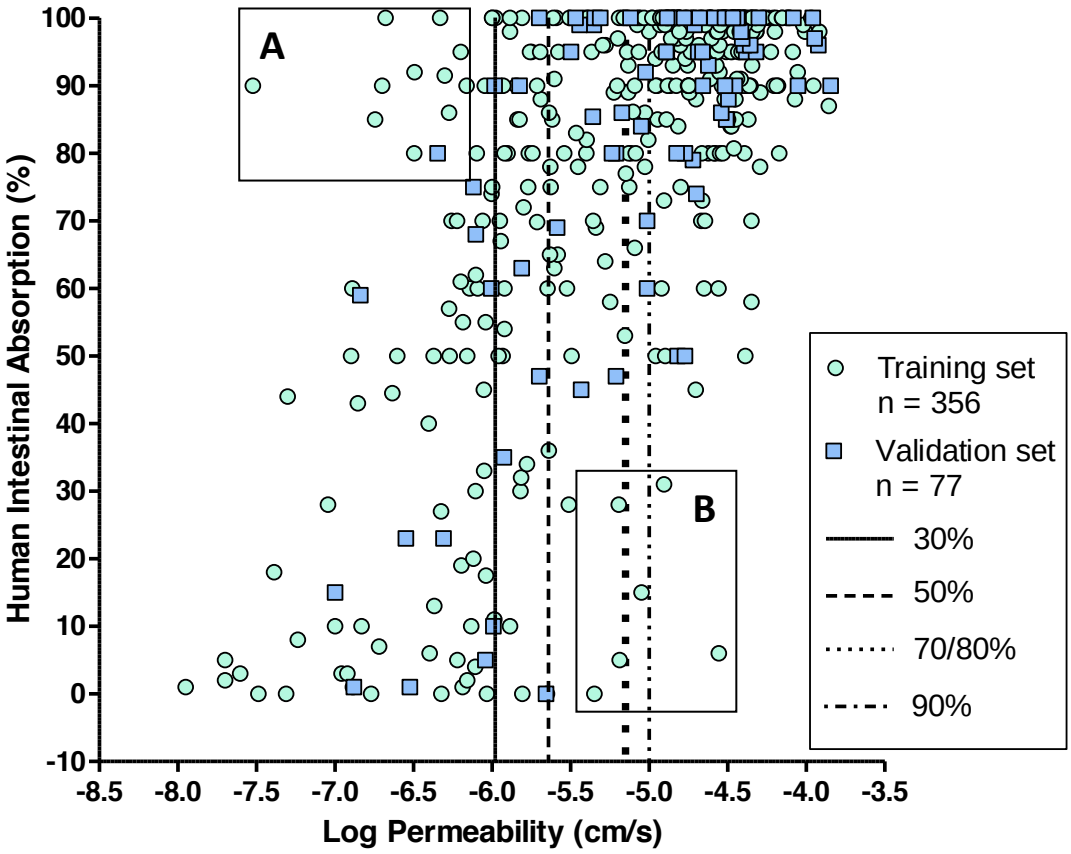

Figure 9.2 Permeability thresholds determined by C&RT analysis with higher misclassification costs applied to false positives for different HIA cut offs of 30%, 50%, 70%, 80% and 90% on %HIA versus permeability plot including areas of pronounced outliers (A= low permeability, high oral absorption; B = high permeability, low oral absorption)

As can be seen in Figure 9.2, there is a correlation between fraction absorbed and permeability. It is common in the literature to assume a sigmoid fit to the relationship between HIA and permeability (Pham-The *et al*., 2013b, Tavelin *et al*., 2003, Varma *et al*., 2012). However, there are too few points at the lower plateau region to justify fitting a sigmoidal fit from a statistical point of view; in spite of this I found a r^2 of 0.435 for a sigmoid fit to the whole 433 compounds. The collection of more data in the 0-50% region may resolve this problem.

From Figure 9.2, there are compounds that are highly-absorbed but have permeability values below the threshold and *vice versa*. The most pronounced outliers have been shown in Figure 9.2 using boxes A and B. Compounds with low

permeability but high fraction absorbed (Region A on Figure 9.2) have been identified as mainly highly soluble and substrates for influx carrier mediated transporters. Examples of these are ribavirin and lamivudine (Shugarts and Benet, 2009, Minuesa *et al*., 2009). Due to the lower levels of these transporters, particularly PEPT1 *in vitro*, the cell permeability underestimates the percentage absorbed of this set of compounds. On the other hand, compounds with high permeability but low fraction absorbed tend to be those that are susceptible to gut metabolism and poorly soluble (Region B on Figure 9.2). Examples of compounds in this outlier group are lovastatin and tacrolimus (Hebert, 1997, Jacobsen *et al*., 1999).

Although the liver is the main metabolising organ, gut metabolism can contribute significantly to overall metabolism and should be considered (Gertz *et al*., 2010). Compounds susceptible to gut metabolism, specifically CYP3A4 substrates, are highly permeable *in vitro* but are poorly-absorbed *in vivo*. However, there are other CYP3A4 substrates in this dataset which do not appear to undergo extensive gut metabolism so they are both highly-absorbed and highly permeable. Reasons for the absorption of some compounds being affected by gut metabolism and others not, even though they are both CYP3A4 substrates, could be the different biotransformation rates by this enzyme, solubility/ dissolution rate of the compound, passive permeation rate, dose amount and substrate affinity (Fagerholm, 2007, Gertz *et al*., 2010, Lin *et al*., 1999). A list of these compounds in regions A and B in Figure 9.2 can be found in the Table A3.4 in Appendix 3.

9.3.3. Oral Absorption Prediction Using Solubility, Dose Number and Melting Point

From Figure 9.2, I have identified potential outliers in the relationship between oral absorption and permeability. Using the models built with permeability and solubility parameters and molecular descriptors, these misclassified compounds could be classified correctly due to the influence of solubility and other related parameters on oral absorption. For example, false positives are highly permeable compounds with poor oral absorption. These compounds may be poorly soluble compounds or those undergoing gut metabolism.

C&RT classification models to predict highly-absorbed or poorly-absorbed class of compounds (HIA ≥ 80 or < 80%) were built using the training sets described in the Methods section 9.2.2. The permeability for ≥ 80% absorption (at -5.15 log scale according to Table 9.5) was used to develop the models. The 80% class definition was chosen as using lower HIA% values to define high or low absorption led to a very low number of poorly-absorbed compounds, compared with highly-absorbed compounds which would seriously reduce significance of models. The HIA 90% cut-off for class definition, although used in some previous work, was not chosen in this work as (based on my preliminary analysis) that definition resulted in poor overall accuracy in the produced models, and the 90% threshold is too high to predict oral absorption class effectively. Selected C&RT models produced for the prediction of HIA class (HIA > or ≤ 80%) using permeability and solubility related parameters and molecular descriptors are shown in Table 9.5. Note that for all models permeability was always used as the first split variable and the table gives the variables used for the second splits. After the second splits, C&RT picks the most significant parameter out of all the molecular descriptors and physicochemical properties available. In Table 9.5, in model 1 after permeability as the first split variable, C&RT automatically builds the rest of the tree by selecting the most significant property/molecular descriptor. For models 2-4, solubility or calculated solubility (GSE method or melting point based absorption potential (MPbAP)) were used on both (high and low permeability) sides of the tree for the second split, and after this C&RT automatically built the rest of the tree. Models 5-10 were built using different combinations of solubility and related parameters on either the high or low permeability side of the trees. Finally, models 11-12 were combinations of the molecular descriptors and solubility related parameters in high or low permeability sides of the trees.

Table 9.5. The results of C&RT analysis for the best permeability and solubility related trees using permeability threshold for ≥ 80% or < 80% HIA as the first split

Model	Parameter used for second split		Misclassification cost ratios (FP:FN)		Dataset	n	SP x SE	SE	SP
	High permeability compounds	Low permeability compounds	High permeability compounds	Low permeability compounds					
1	Molecular Descriptors[a]	Molecular Descriptors[a]	3:1	6:1	t	356	0.720	0.754	0.955
					v	77	0.519	0.593	0.875
2	Solubility (mg/ml)	Solubility (mg/ml)	2:1	10:1	t	241	0.723	0.823	0.879
					v	54	0.618	0.674	0.917
3	GSE solubility	GSE solubility	2:1	1:1	t	261	0.695	0.891	0.779
					v	53	0.638	0.829	0.769
4	MPbAP	MPbAP	1:1	1:1	t	249	0.753	0.876	0.859
					v	48	0.631	0.757	0.833
5	Solubility (mg/ml)	GSE solubility	2:1	10:1	t	200	0.754	0.820	0.920
					v	40	0.583	0.667	0.875
6	Dose number	MPbAP	2:1	10:1	t	196	0.758	0.791	0.958
					v	40	0.636	0.636	1.000
7	MPbAP	GSE solubility	2:1	1:1	t	256	0.723	0.884	0.818
					v	51	0.667	0.800	0.833
8	MPbAP	Solubility (M)	2:1	1:1	t	197	0.776	0.866	0.896
					v	40	0.697	0.697	1.000
9	Solubility (mg/ml)	Solubility (M)	2:1	10:1	t	241	0.754	0.766	0.985
					v	54	0.533	0.581	0.917
10	GSE solubility	Solubility (M)	2:1	1:1	t	201	0.722	0.881	0.82
					v	40	0.663	0.758	0.875
11	GSE solubility	Molecular Descriptors[a]	2:1	1:1	t	262	0.717	0.887	0.809
					v	53	0.650	0.780	0.833
12	MPbAP	Molecular Descriptors[a]	2:1	1:1	t	257	0.746	0.880	0.848
					v	51	0.688	0.750	0.917

[a] These are the molecular descriptors statistically selected by C&RT out of all the molecular descriptors and solubility parameters. FP: false positive; FN: false negative; GSE: General solubility equation; MPbAP: melting point based absorption potential. SP x SE, Overall Accuracy; SE, Sensitivity; SP, Specificity

From Table 9.5, it is interesting to note which properties were used to build the selected models. Note that many combinations of melting point, dose and solubility related parameters were tested and Table 9.5 is a selection of the best models based on accuracy (SE X SP). Using melting point did not yield high prediction models (data not shown). It was thought that due to the relationship between melting point and solubility this parameter might be a useful alternative to solubility, as these two properties share similar functions such as enthalpy energies which must be overcome in order to solubilise or melt. Additionally, dose number was useful only for splitting the high permeability compounds and the combination with MPbAP yielded a good prediction model (Model 6 in Table 9.5). Dose number is used to define high and low solubility for the BCS system (Amidon *et al*., 1995, CDER/FDA, 2000). By definition, increasing the dose or a low solubility will result in a high dose number and this is expected to lead to poor oral absorption of highly permeable compounds.

The majority of the selected models in Table 9.5 incorporate solubility and predicted solubility especially for highly permeable compounds. Unlike GSE solubility which was used on both sides of the C&RT trees, MPbAP only yielded good models when used for splitting on the high permeability compounds. Experimental solubility in two units, mg/ml or molar, has been used in models. Solubility in M, which takes into account the molecular weight and is smaller for high molecular weight compounds, was successful when utilised for splitting of the low permeability compounds (Models 8, 9 and 10).

In terms of the role of solubility in the absorption process, one would expect poor absorption of poorly soluble compounds, due to solubility being the rate limiting factor in absorption. However, this is not the picture presented by the classification trees 1-12 (See Appendix 3, Figures A3.8-A3.19). According to the classification tree models, the low permeability and high solubility compounds always have low intestinal absorption (< 80%). This is probably due to the highly polar nature of such compounds. On the other hand, poorly water soluble compounds of low permeability may be highly-absorbed from the small intestine if they have small polar surface area (models 3-7) or a low small sum of absolute atomic partial charge, ABSQ (models 2, 8, 9, 10), which also indicates polarity of molecules. The absorption limiting effect of poor aqueous solubility is not seen for highly permeable compounds either. Here, highly permeable compounds with poor aqueous solubility are still highly absorbable

from GI, with the exception of compounds with high polar surface area, low dipole moment (models 2, 5, 9) or small Balaban Topological index which is an indicator of molecular shape (models 3, 4, 10, 11). The reason for not observing the limiting effect of poor aqueous solubility here could be, firstly, the lack of enough representation of these solubility limiting compounds in the dataset, and secondly, the effect of formulation of oral dosage forms with measures taken for improved dissolution rate (excipients, particle size, etc.), which could mask previous solubility limiting effects of such compounds.

The top molecular descriptors used in models 1-12 in Table 9.6 are PSA and Balaban topological index. Both of these descriptors are related to both absorption and solubility prediction models (Clark, 1999, Bergstrom, 2005). PSA has been described in previous chapters. The Balaban topological index, *J*, is the average-distance sum connectivity and relates to the shape of the molecule (Balaban, 1982). The next popular descriptors are sum of absolute charges on each atom of the molecule (ABSQ) (Gasteiger and Marsili, 1980) and lowest unoccupied molecular orbital energy (LUMO) calculated by VAMP.

9.3.4 Selected C&RT Models

In order to generally compare models 1-12 from Table 9.6, the compound datasets used to build the resulting models should be taken into account. The degree of difficulty of the classification model will change depending on the compounds in the dataset. The model with the highest SP x SE for the validation set is model 8, with a value of 0.697. However, this is based on a training set of only 197 and a validation set of 40 compounds, due to the missing experimental solubility or melting point values. On the other hand, model 12 has a slightly lower SP x SE of 0.682 for the validation set, but it was built using a training set of 257 and assessed using a validation set of 51 compounds. Therefore this model compared to model 8 covers a wider chemical space so will be able to predict for more compounds without extrapolation (higher generalization ability to new compounds). Moreover, the only experimental parameter used in this model is melting point, which is used for the calculation of MPbAP. I also selected model 7, which has used calculated solubility and MPbAP, and model 3, which has used only the calculated solubility to indicate

the roles of solubility and absorption potential. The corresponding C&RT models are presented in Figures 9.3-9.5.

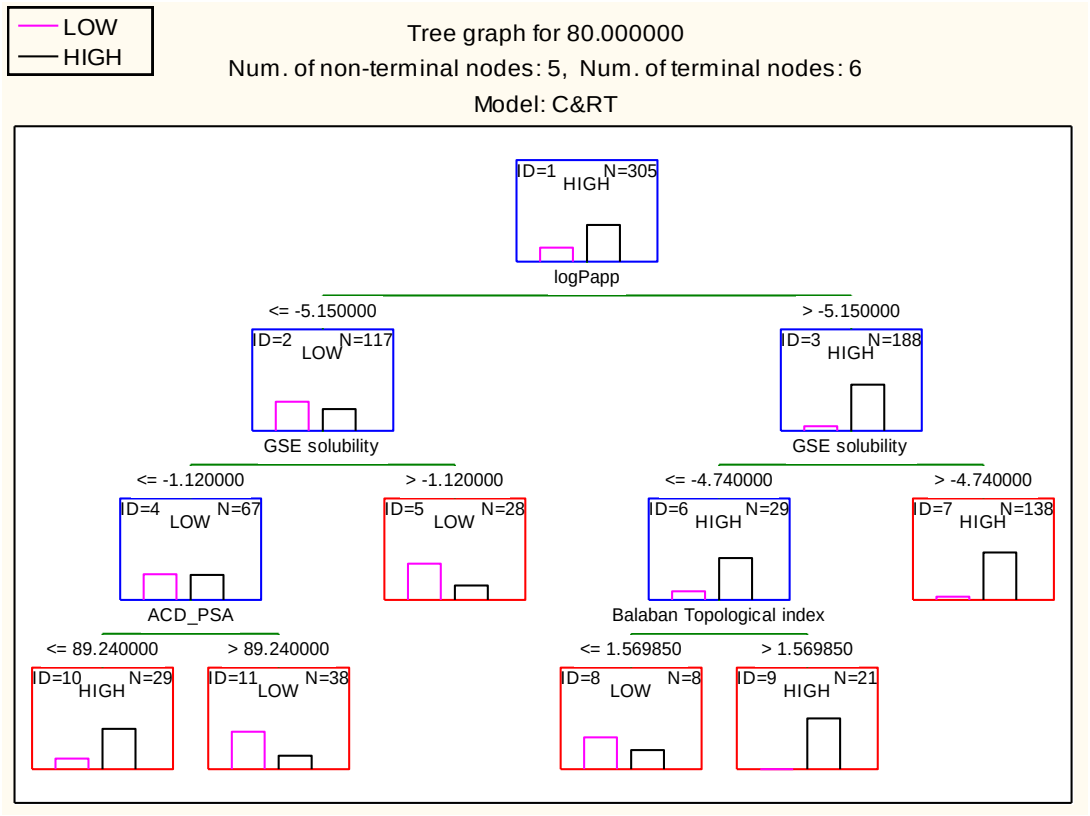

Figure 9.3. Model 3 - C&RT permeability and predicted solubility (GSE) model when higher misclassification costs of two to reduce false positives were applied to high GSE solubility node

In Figure 9.3, Model 3, permeability is used as the first C&RT split variable and then calculated solubility from GSE equation on both sides of the tree is used as the second split variable. Polar surface area and Balaban index were picked automatically by the C&RT analysis. The model shows that highly permeable and highly soluble compounds have high intestinal absorption (node 7). Moreover, compounds with low predicted solubility ($\leq$ -4.74) can still be classed as highly-absorbed if the Balaban index is > 1.57. Compounds with a low Balaban index will be poorly-absorbed and such examples include mebendazole and ketoconazole. In spite of this, there are misclassifications in this node 8 in Figure 9.3; ziprasidone and tiagabine are misclassified as poorly-absorbed when in fact they have HIA $\geq$ 80%. Balaban topological index, J, a highly discriminant topological descriptor, gives an indication of shape including branching and cyclicity of a molecule. A high index

can indicate a high number of branches, close proximity of the position of these branches, as well as increased number of double bonds on a molecule. A low index can indicate a low level of branching as well as a larger number of cyclic groups (Balaban, 1982). The relationship between Balaban index and solubility with reference to melting point has been shown previously in the literature (Ghafourian and Bozorgi, 2010). In spite of this, there is not much difference between the calculated GSE solubilities between the two nodes although there is a significant difference between the average melting points (222 ^{o}C compared with 193 ^{o}C in nodes 8 and 9 respectively), suggesting a possible effect of melting point on absorption.

Poorly permeable compounds can be highly-absorbed only for compounds with predicted solubility ≤ -1.12 log unit if the PSA is low. This is a higher solubility value than the threshold seen in splitting of node 3, and is not expected to limit the intestinal absorption. There are some misclassified compounds in this group, which are actually poorly-absorbed despite having a low PSA, but are classified as highly-absorbed according to this tree. The reasons for misclassifications are mostly due to efflux mechanisms reducing the absorption of compounds. Examples include nadolol and norfloxacin, which both have low PSA and are classed as highly-absorbed but are observed to have poor oral absorption due to transporter effects (Matsson *et al*., 2005, Merino *et al*., 2006). Unlike nadolol, which is classed as highly soluble, norfloxacin is considered as a poorly soluble compound in class IV of the BCS system. One may speculate that the presence of more such compounds in this dataset may have led to further split of this node based on lower solubility cut-offs to class compounds with extremely low aqueous solubility as poorly soluble.

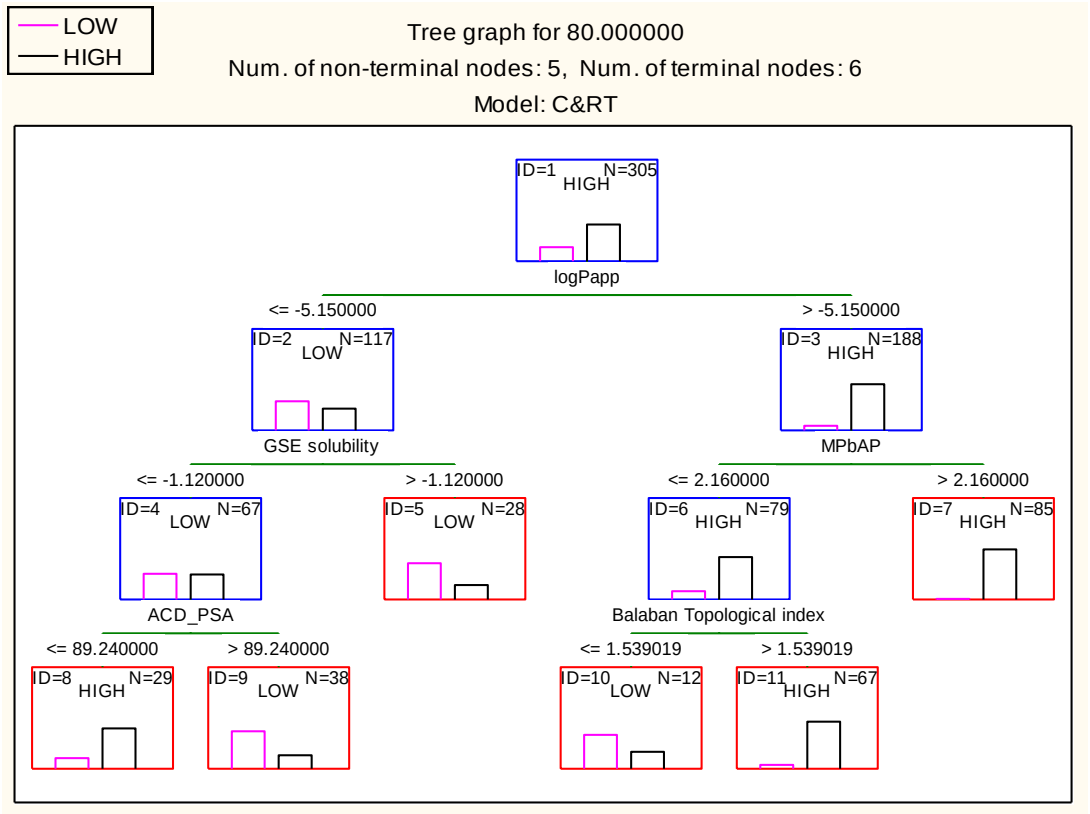

Figure 9.4. Model 7 - C&RT permeability, predicted solubility (GSE) and MPbAP model when higher misclassification costs of two to reduce false positives were applied to GSE node

Model 7 was built using GSE solubility for the second split of the poorly permeable compounds (node 2) and MPbAP for the second split of highly permeable compounds in node 3. This model was chosen due to high validation SP x SE using a larger training and validation sets. The descriptors used in this tree are the same as in Figure 9.3. Model 3, however, using the split based on MPbAP, appears to split more compounds into node 6 to be classed by Balaban topological index. In this tree a lower threshold of 1.54 for Balaban topological index increases the number of correctly classified poorly-absorbed compounds when permeability is high. Examples of this type of compounds include the BCS class II compounds spironolactone and ketoconazole.

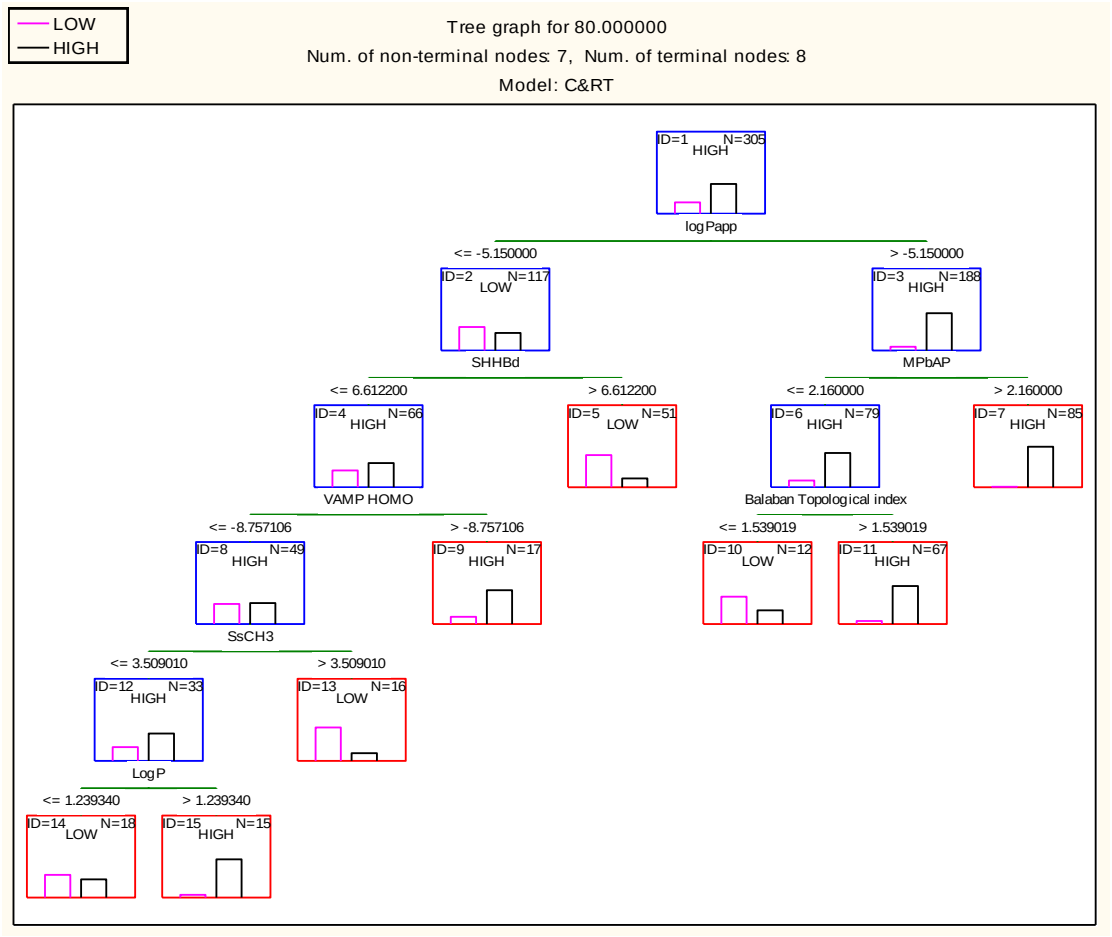

Figure 9.5. Model 12 - C&RT permeability and MPbAP model when higher misclassification costs of two to reduce false positives were applied to permeability node

From Figure 9.5, classification of highly permeable compounds in node 3 is the same as Figure 9.4. Poorly permeable compounds with a high number of hydrogen bonding donors (SHHBd >6.61) will be poorly-absorbed, which is confirmed by the literature such as Lipinski's rule of five (Lipinski *et al*., 1997), which has been explained in previous sections of this thesis. Compounds can be misclassified as poorly-absorbed based on a higher number of hydrogen bond donor groups mainly due to being highly-absorbed due to substrate specificity for influx transporters. Examples of misclassified compounds include ribavirin and folinic acid.

A poorly permeable compound will still be highly-absorbed if HOMO energy is greater than -8.76. A comparison of the molecular structures in this node indicates that these compounds have more aromatic rings compared with compounds with

lower HOMO energy (node ID 8) where the average number of aromatic rings is one. In addition, it was also found that a number of low HOMO compounds had a permanent quaternary ammonium or ionisable centre, such as trospium and neostigmine.

Even if a poorly permeable compound has a low HOMO energy, it can still be classed as highly-absorbed if the compound has few methyl groups (SsCH3 ≤ 3.509) or log P > 1.239. Compounds with logP < 1.24 are classified as poorly-absorbed, but there are false negatives such as orally administered cephradine and baclofen, which are both highly-absorbed but are predicted as poorly-absorbed by having a low logP. The reason for some of the false negatives in this node is that some of these compounds are substrates for influx carrier mediated systems.

9.3.5 Discussion of Related Literature

9.3.5.1 Subjective Definition of a Permeability Threshold for Oral Absorption Prediction

Permeability from *in vitro* cell based assays has been utilised frequently in the literature. These thresholds are then used to give an indication of potential oral absorption from permeability data. A summary of a few permeability thresholds defined by other works is shown earlier in this thesis (Chapter 4, Table 4.1)

Early permeability thresholds in the literature are commonly based on small compound datasets. Artursson *et al* (Artursson and Karlsson, 1991) set a permeability threshold of > 1 x 10^{-6} for complete absorption based on 20 compounds. Based on other works in the literature this value is too low to predict complete absorption, whereas other works have permeability thresholds one order of magnitude higher. For example, Yee *et al* (Yee, 1997) has stated that a threshold > 10 x 10^{-6} permeability is related to absorption > 70%. What is apparent is the difference between permeability thresholds from different sources, which is dependent on the small number of compounds tested and inter and intra laboratory differences (Hou *et al*., 2007c). In comparison, my permeability thresholds are statistically defined by C&RT rather than a subjective determination; the thresholds picked by C&RT are similar to those in the literature, especially when high

absorption was set at either as > 70%, > 80% or > 90%, indicating that high absorption is related to permeability > 7 x 10^{-6} cm/s. The permeability threshold determined by Hou *et al* (Hou *et al*., 2007c) of 6 x 10^{-6} cm/s is based on data from numerous sources and is very similar to my 70 - 90% class permeability thresholds.

Di *et al* (2011) (Di *et al*., 2011) used MDCK II cells with low efflux endogenous transporter expression (MDCK-LE) to define a threshold of 3 x 10^{-6} to distinguish between low/medium absorbed compounds (< 80% HIA) and highly-absorbed compounds. A dataset published by Varma *et al* (Varma *et al*., 2012) using the MDCK-LE cell line shows that the permeability threshold defined by ROC analysis using this cell line (≥ 5.0 x 10^{-6} cm/s) is similar to Caco-2 thresholds in the literature. This value is in agreement with C&RT permeability thresholds in this work. The threshold similarity between Caco-2 and MDCK cell lines is expected by the linear relationship between these two cell lines shown in this work.

Finally, more recently Pham-The *et al* (2013) (Pham-The *et al*., 2013b) established a rank order relationship between Caco-2 permeability and oral absorption for 324 compounds. The thresholds defined were based on standard compounds from the FDA with known fraction absorbed values. For example, for a compound to be considered highly-absorbed, it must have an apparent permeability greater than metoprolol, a FDA standard compound with known HIA. In this case, Caco-2 permeability greater than 16 x 10^{-6} cm/s, which is 0.8 times the metoprolol permeability, was used to take into account the lower HIA threshold of 85% used. For the low absorption threshold, an average value of 0.7 x 10^{-6} cm/s, based on the permeability of mannitol, was used. In this study this threshold was used to define compounds with HIA < 30%. However, mannitol has a reported HIA of ~18%, therefore the use of this permeability threshold may increase the number of false negatives.

9.3.5.2 The Influence of Permeability and Solubility on Oral Absorption Modelling

Permeability and solubility are two important factors important for oral absorption. Therefore, the effect these two properties have on oral absorption and in turn how they influence oral absorption prediction is important to establish. From the literature, there is a lot of focus on permeability, and as shown in this work there is a

rank order relationship between HIA and permeability. On the other hand, solubility seems not to be regarded as important as permeability in relation to oral absorption, but as a factor that can lead to poor (solubility limited) absorption in addition to other limiting factors, such as transporter and enzyme effects. Furthermore, the relative importance of solubility could be dependent on the research organization and the mechanistic importance of solubility in regards to oral absorption may not be considered (Lipinski, 2000). In spite of this, the main reasons for poor oral absorption have been shown to be either poor permeability or poor solubility or both (Savjani *et al*., 2012).

The results of this work indicate that permeability is the most important parameter influencing oral absorption prediction. Permeability was always picked as the top molecular descriptor when building C&RT models. In contrast, solubility and the related parameters were never picked as the top descriptor or even in the second split, unless selected manually at this second level in order to examine if there was any influence of solubility on oral absorption prediction.

It is apparent that solubility can be a rate-limiting step in oral absorption (Zhao *et al*., 2002, Sugano, 2011, Amidon *et al*., 1995). This is based on the principle that a drug must be dissolved in the gastrointestinal fluid in order to then permeate the membrane to be absorbed. However, formulation development strategies can overcome this problem, for example by employing solubilising agents, pH control, or complexation (Stegemann *et al*., 2007).

In any case, the results obtained here do not directly indicate the poor absorption of poorly soluble compounds and the effects of poor solubility in limiting absorption. According to this chapter, in general, compounds that are highly permeable but have low solubility can be predicted as highly or poorly-absorbed depending on the other molecular properties. Moreover, poorly permeable but highly soluble compounds are classed as poorly-absorbed, although there are exceptions to this, i.e. the false negatives. One important consideration in analysing these results is the threshold of solubility in the models. For example, poorly permeable compounds with poor solubility may have high oral absorption (see models 3 and 7 for example). However, it must be noted here that poor solubility has been defined as < -1.12 log

unit, which is quite high when comparing with the threshold values suggested in the literature for BCS classes II and IV (Amidon *et al*., 1995). A further observation from the models could be the poor representation of very poorly soluble compounds in the dataset, i.e. those having solubility-limited absorption. As a result, it may not be statistically advantageous to further split the classification tree to allocate these compounds into a separate terminal node. For example in a large dataset of fraction absorbed, 24 were highlighted to have solubility issues out of 647 compounds (Hou *et al*., 2007c). Besides this, the formulation techniques may improve the dissolution rate of these compounds and overcome the low solubility issues of compounds in the fraction absorbed dataset used in this work.

It is difficult to directly compare other models in the literature with this work, as different datasets and methods have been used. Early oral absorption models which use a diverse dataset are too small to represent all the different biological processes of absorption and other factors such as solubility. The majority of oral absorption models in the literature do not include compounds which have solubility issues (Wessel *et al*., 1998, Bai *et al*., 2004). Therefore, these and other models may only be useful for predicting absorption for compounds with no solubility issues. In addition, some of these studies also removed compounds with transporter effects or compounds with a permanent charge (Egan *et al*., 2000, Hou *et al*., 2007c). This simplifies the resulting models by removing the compounds with rate-limiting steps. However, the main issue with this is the potential impact on the generalizability of the resulting models, which will fail to predict the oral absorption of these excluded compound classes, despite the increased need in current drug discovery projects for prediction of absorption of the increasingly poorly-soluble compounds.

In studies by Zhao and co-workers, data with poor solubility and dose dependency were highlighted and not used in the majority of the initial models. When these compounds were included, the resulting models had higher error (Zhao *et al*., 2001). It was also noted, however, that the more insoluble a compound the lower the resulting absorption. In a later study by Zhao, compounds identified with no solubility issues were used to build models and some of these resulting models were then used to predict absorption for the compounds with dose-limiting and dose dependency effects. Overall prediction of absorption of these excluded compounds

was in agreement with observed values or the models tended to overestimate absorption (Zhao *et al*., 2002). My oral absorption models are able to predict oral absorption class even for majority of compounds with with poor solubility, by incorporating molecular descriptors in addition to permeability and solubility into the models. From the list of 27 compounds with solubility and related problems defined by Zhao *et al* (Zhao *et al*., 2002), 14 were utilised in this work with experimental permeability and solubility values present. Using the best models chosen, 11 out of 14 compounds were predicted correctly by model 3, 12 out of 14 correct predictions by model 7 and all 14 compounds were predicted correctly using model 12.

With the extended use of BCS classification in drug discovery, the influence of solubility and permeability is of great interest (Pham-The *et al*., 2013a). In work by Pham-The *et al* (2013), oral absorption was predicted taking into account solubility, which is a general aim of the BCS. In this study, Pham-The, using a rank order relationship, noted that the relationship between permeability and oral absorption is less certain for poorly-absorbed compounds, which is a similar observation to my results. They also found various contour plots showing that incorporating solubility improves classification of HIA based on permeability data by about 10%; therefore showing that using solubility in models is potentially advantageous for oral absorption prediction.

From the literature examples, as well as this work, the influence of solubility could be included to help predict oral absorption. However, the main issue is the lack of experimental solubility for drug compounds to be used in oral absorption modelling. The use of experimental solubility data in the prediction of oral absorption alongside permeability yields good accuracy to predict oral absorption; however, the lack of experimental solubility limits the application for the prediction of new compounds. Therefore, according to my results, predicted solubility descriptors such as GSE solubility and parameters such as MPbAP can be used successfully instead of experimental solubility. These are based on simple properties of lipophilicity, melting point and dose. Despite this, melting point alone was not successful in providing an adequate alternative to experimental solubility, even though partition coefficient was also available to be used concurrently in the same model. Due to the

complexity of solubility, it is difficult to find one molecular descriptor to adequately describe all the solubility processes.

9.4 Conclusion

The two main properties influencing oral absorption are permeability and solubility. In order to establish the relationship of these two properties with oral absorption classification, firstly, a larger dataset was established from different sources. This was made possible through combining Caco-2 and MDCK permeability after verifying that there is a linear relationship between these two cell lines, even for compounds with different absorption mechanisms.

Secondly, using the combined permeability dataset, permeability thresholds for various levels of oral absorption were investigated using C&RT analysis. Due to the larger number of highly-absorbed compounds, misclassification costs were applied and improved the threshold definitions statistically. The thresholds obtained from the objective C&RT analysis are similar to some of those in the literature using mainly subjective methods to determine permeability thresholds.

Finally, the permeability thresholds were then used to build decision trees with the C&RT method, incorporating solubility and related parameters, as well as the calculated molecular descriptors to predict oral absorption class. Melting point is not a useful parameter to predict absorption when used stand-alone. However, when melting point is utilised to calculate combined parameters such as predicted (GSE) solubility and melting point-based absorption potential, it yielded high accuracy models compared with experimental solubility. This is due to the possibility of using more data for the training of the models when calculated or more easily accessible experimental parameters are used. Therefore, models built using predicted values of solubility and melting point-based absorption gave rise to better predictive models. Molecular descriptors utilised in the models, such as those describing size, shape, polarizability and hydrogen bonding, can be related to both permeability and solubility, and therefore to oral absorption. These molecular descriptors were shown to be necessary for oral absorption models, in order to correctly classify the compounds with solubility-limited absorption. The models built in this work are

useful for a better mechanistic understanding of the effect of these properties and how they contribute to overall oral absorption.

10 Comparing Two Multi-Label Classification Methods for Provisional Biopharmaceutics Classification System (BCS) Class Prediction

10.1 Introduction

The Biopharmaceutics Classification System (BCS) classifies orally administered compounds into one of four classes based on their permeability, solubility and related properties (Amidon *et al*., 1995, CDER/FDA, 2000). The use of the BCS in drug discovery and development can help streamline chemical/formulation optimisation to improve solubility and/or permeability (Ku, 2008, Varma *et al*., 2012, Pham-The *et al*., 2013a, Lennernas and Abrahamsson, 2005, Butler and Dressman, 2010, Bergstrom *et al*., 2003).

The majority of classification algorithms in the literature carry out single label classification to create separate models of oral absorption, solubility or permeability. However, single label classification cannot take into account the interactions between permeability and solubility, therefore multi-label classification can be carried out in order to take such interactions into account (Carvalho and Freitas, 2009, Tsoumakas and Katakis, 2007, Read *et al*., 2011).

There are many works in the literature that assign BCS for drug compounds (Lindenberg *et al*., 2004, Takagi *et al*., 2006, Dahan *et al*., 2009). In spite of this, there are relatively few studies that utilise computational models to predict BCS class in a multi-label fashion using large datasets. Therefore, the aim of this work is to compare two multi-label methods, namely binary relevance and classifier chain, for the prediction of BCS using permeability and solubility as classes. To my best knowledge there are no other works in the literature which compare multi-label methods for provisional BCS prediction suitable for use in drug discovery. Binary relevance is a simple multi-label method; however it has the disadvantage that it cannot take into account any interactions between the labels. Based on this, this chapter introduces the classifier chain multi-label classification method to the area of pharmacokinetics – to the best of my knowledge, this is the first work using

classifier chains in pharmaceutical sciences. It is anticipated that, by using this method and taking into account the label interactions, more accurate models can be produced for provisional BCS prediction. This chapter shows the potential of multi-label classification methods, which can be used for the future prediction of many pharmacokinetic properties in drug discovery and development.

10.2 Methods

10.2.1 Dataset

Dataset 4 (as described in the Datasets and Methods section 5.1.4) was utilised for the prediction of BCS class as defined in the main method section of this thesis.

Based on previous chapter 9, the benchmark threshold to define the boundary between high and low permeability for 80% Human Intestinal Absorption (HIA) was set at 7.08 x 10^{-6} cm/s (logPapp of -5.15). Therefore, a compound with *in vitro* permeability < 7.08 x 10^{-6} cm/s would be defined as poorly permeable and a compound with permeability ≥ 7.08 x 10^{-6} cm/s would be defined as highly permeable.

In the BCS, the definition of the boundary between high and low solubility is determined using the dose number ($D_o = (M_o/V_o)/S$), where M_o is the highest dose strength, V_o is 250ml and S is the aqueous solubility (mg/ml)), compounds with $D_o \leq 1$ are classed as highly soluble and drugs with $D_o > 1$ are assigned as poorly soluble drugs (Amidon *et al*., 1995, CDER/FDA, 2000). However, in early drug discovery the clinical dose is usually unknown; therefore a suitable threshold needs to be defined. Additionally, D_o is a property of the drug formulation and not a specific property of the active compound. For this chapter, a solubility cut off of 0.2 mg/ml was set. Hence, any drug with solubility ≥ 0.2 mg/ml was defined as highly soluble and drugs with solubility < 0.2 mg/ml were classed as poorly soluble. A value of 0.2 mg/ml was used as, according to Lipinski *et al*. (Lipinski, 2000), this value is the minimum solubility required to get a projected clinical dose of 1 mg/kg for compounds with low permeability. This cut-off for solubility has also been used in a recent work for BCS using MDCK permeability and solubility (Varma *et al*., 2012).

10.2.2 Training Sets and Validation Sets

The compounds in each permeability and solubility dataset were sorted based on either ascending logPapp or logS (mg/mL) separately (excluding the 127 compounds used for the external BCS validation set). For each individual dataset, from each group of five consecutive compounds, four were assigned to the training set, and one compound was allocated to the validation set randomly. By doing this, a similar distribution of values in the training and validation sets was achieved for both datasets. The resulting compound numbers in the training and validation sets are shown in Table 10.1.

Table 10.1. Training and validation set compound numbers used in chapter 10

Type of dataset	Training n	Validation n	BCS validation n
Permeability	1026	262	127
Solubility	490	133	127

The training sets were used to build separate models to predict permeability and solubility classes. The individual validation sets for the permeability and solubility datasets were used to measure the predictive performance of the individual models for the two types of classes. Lastly, in order to compare the two multi-label methods for provisional BCS classification, an additional external validation set containing 127 compounds with known permeability and solubility values was used (BCS validation set).

10.2.2 Model Development

10.2.3.1 Molecular Descriptors

Molecular descriptors were calculated using TSAR 3D v3.3 (Accelrys Inc.), MDL QSAR (Accelrys Inc.), MOE (Chemical Computing Group Inc.) v2012.10 and Advanced Chemistry Development ACD Laboratories/LogD Suite v12. A total of 492 molecular descriptors were generated and made available to the classification algorithm before feature selection.

10.2.3.2 Feature Selection

Firstly, molecular descriptors with more than 10 missing values were removed, so that 14 molecular descriptors were removed from each training set and this resulted in 478 molecular descriptors available for pre-processing feature selection.

Based on chapter 8, I used the predictor importance ranking in random forest to obtain the top 20 molecular descriptors. (See section 5.4.4 for explanation of this method.) Using only the training set, optimisation of the random forest was carried out based on the plot of misclassification rate vs the number of trees. Based on this plot, the optimum number of trees was selected (106 for the solubility, 109 for the permeability). The maximum number of levels for each tree was set to the default 10. The top 20 molecular descriptors for each property (solubility and permeability) can be found in Appendix 4 (Tables A4.1 and A4.2).

10.2.3.3 QSAR Modelling Techniques

STATISTICA v12 (StatSoft Ltd.) software was used for building each classification model using C&RT analysis. C&RT analysis is a statistical technique that uses decision trees to solve regression and classification problems developed by Breinman *et al*. (Breiman *et al*., 1984).

For the binary relevance method, each class – i.e. solubility or permeability variable – was set as the dependent variable and binary classification was carried out using selected molecular descriptors as the independent variables to create individual models for each class label.

For the classifier chain method, initially individual solubility classification models were built using the top 20 molecular descriptors as chosen by feature selection. These models were then used to predict the solubility class for the whole permeability dataset. The permeability model was then built by setting permeability class as the dependent variable, while the predicted solubility and the top 20 molecular descriptors pre-selected for permeability were set as the independent variables. The preliminary results indicated that predicted solubility class (acting as a molecular descriptor) would not be used high up in the tree (if at all); therefore predicted solubility was selected manually as the first molecular descriptor in the

C&RT model for permeability. The rest of the C&RT decision tree was allowed to be built automatically.

For this chapter, the stopping factors used when growing the C&RT tree were the minimum number of compounds for splitting. These stopping factors were the default values for the software and are based on the number of compounds in the dataset. This enables pruning of the tree and prevents over-fitting of the decision tree. For the permeability and solubility datasets, stopping factors of 25 and 12 respectively were used.

10.2.3.4 Misclassification Costs for Classification Models

As shown in chapters 7-9, misclassification costs are a useful method to overcome the dataset bias of unbalanced class distributions (where one class value is much more frequent than another) without reducing dataset size (Newby *et al*., 2013a, Newby *et al*., 2013b). Even if the dataset has a balanced class distribution, the application of higher misclassification cost for a specific class can increase the predictive accuracy and reduce misclassification errors of that specific class.

The solubility and permeability datasets have roughly balanced class distributions, therefore misclassification costs can remain as equal (FP:FN of 1:1, where FP:FN is the ratio of the number of false positives to the number of false negatives). However, usually there is an under-representation of BCS classes 3 and 4 due to the low number of poorly permeable compounds and compounds with both poor permeability and poor solubility. Therefore, in order to potentially improve the predictive accuracy of these under-represented classes, higher misclassification costs can be applied to reduce false positives (i.e. the number of compounds in the poor solubility and poor permeability classes which are wrongly predicted as having high solubility or high permeability), in order to take into account the lack of compound representation for these classes when combining the solubility and permeability predictions. A higher misclassification cost of 1.5 was applied to the false positive class (FP:FN of 1.5:1) based on the data distribution of the permeability and solubility datasets.

10.3 Results

10.3.1 Permeability and Solubility C&RT Models

In this chapter, I have investigating the use of two multi-label classification methods to predict provisional BCS class using permeability and solubility from the literature and published datasets. Separate models of permeability and solubility were built using training sets of 1026 and 490 compounds respectively, using the top 20 molecular descriptors selected by the random forest-based feature selection method. The predictions from the solubility and permeability models were then combined to give a provisional BCS class for an BCS validation set of 127 compounds. All the C&RT decision trees that produced the results reported in Tables 10.2 and 10.3 in this chapter can be found on the accompanying disk with this thesis. In Tables 10.2 and 10.3, the best models are those that have the highest SP, SE and SP X SE and the lowest CNMI. These have been highlighted in **bold** for the training and validation sets in these tables. Firstly, the two solubility models whose results are shown in Table 10.2 are models with equal and higher misclassification costs applied to reduce false positives – models 1 and 2, respectively. The compound numbers in training and validation sets for solubility and permeability for Tables 10.2 and 10.3 are lower than the original numbers in Table 10.1. This is because for certain compounds molecular descriptors could not to be calculated and therefore could not be classified in the terminal nodes. Therefore, the compound numbers in Tables 10.2 and 10.3 represent the compound numbers classified by the models.

Table 10.2. Results of C&RT Analysis for the Classification of Solubility

Model	Misclassification cost ratio (FP:FN)	Set	n	SP X SE	SE	SP	CNMI
1	1:1	t	485	0.621	**0.784**	0.792	0.212
		v	128	**0.578**	**0.795**	0.727	**0.234**
2	1.5:1	t	485	**0.638**	0.706	**0.903**	**0.178**
		v	128	0.538	0.658	**0.818**	0.243

t-training; v-validation; Sensitivity is equivalent to the number of correctly classified highly-absorbed compounds and is calculated using SE=(TP/(TP+FN)); Specificity is equivalent to the number of correctly classified poorly-absorbed compounds and is calculated using SP=(TN/(TN+FP)); TP-true positive; FN-False negative; TN-true negative; FP-false positive; Overall Accuracy of the models was calculated by multiplying Specificity by Sensitivity (SP x SE); n, is the number of compounds that was predicted by the model for the training and validation set; CNMI = Cost normalised misclassification index;

Both solubility models from Table 10.2 can be considered the best depending on the intended use and purpose of the model. Model 1 has the highest sensitivity for the training set and validation set as well as overall accuracy for the validation set. Whereas model 2, as expected, has the highest SP for the training and validation set due to the application of higher misclassification costs to reduce false positives. Therefore, if the aim of the model is to predict poorly soluble compounds, model 2 would be the best model; but model 1 would be the best to use if the aim was to predict highly soluble compounds. Model 1 may be considered as the best C&RT model in this work (shown in Figure 10.1), since for the validation set, there is more of a balanced prediction for poorly and highly soluble compounds (higher SP X SE). Both solubility models were then used to predict solubility for compounds in the permeability dataset, which was in turn used as an additional descriptor (independent variable or feature) for building permeability model – this process implements the classifier chain approach for multi-label classification, discussed earlier.

The statistical parameters of the permeability models produced in this work are shown in Table 10.3. Initially, permeability models were built using only the top 20 molecular descriptors selected by the random forest-based feature selection method (models 1 and 4). Next, permeability models were built using the predicted solubility either from the solubility model 1 or from solubility model 2 in Table 10.3 in addition to the top 20 molecular descriptors as the independent variables. Again models were also built with equal (models 1-3) or higher misclassification costs (models 4-6) applied to reduce false positives (FP:FN 1.5:1).

Table 10.3. Results of C&RT Analysis for the Classification of Permeability (with and without predicted solubility incorporated in the model)

Model	Misclassification cost ratio (FP:FN)	Solubility Model included	Set	n	SP X SE	SE	SP	CNMI
1	1:1	none	t	1016	0.653	**0.847**	0.771	0.192
			v	261	0.503	0.727	0.692	0.291
2		1	t	1016	0.655	0.841	0.778	0.191
			v	261	**0.519**	**0.742**	0.699	0.280
3		2	t	1016	0.638	0.761	0.838	0.200
			v	261	0.482	0.641	0.752	0.303
4	1.5:1	none	t	1016	**0.659**	0.807	0.817	0.188
			v	261	0.484	0.664	0.729	0.298
5		1	t	1016	0.630	0.716	0.880	**0.185**
			v	261	0.489	0.586	**0.835**	**0.265**
6		2	t	1016	0.625	0.706	**0.884**	0.187
			v	261	0.489	0.586	**0.835**	**0.265**

t-training; v-validation; Sensitivity is equivalent to the number of correctly classified highly-absorbed compounds and is calculated using SE=(TP/(TP+FN)); Specificity is equivalent to the number of correctly classified poorly-absorbed compounds and is calculated using SP=(TN/(TN+FP)); TP-true positive; FN-False negative; TN-true negative; FP-false positive; Overall Accuracy of the models was calculated by multiplying Specificity by Sensitivity (SP x SE); n, is the number of compounds that was predicted by the model for the training and validation set; CNMI = Cost normalised misclassification index;

Based on the validation set, the best permeability model to choose would be model 2. This permeability model was built using the predicted solubility from model 1 in Table 10.2 and equal misclassification costs applied. This model achieved the highest overall accuracy (SP X SE) and sensitivity for the validation set of 0.519 and 0.742, respectively. In addition, it also had the second highest SP X SE and SE for the training set and the lowest CNMI for the training and validation sets, when comparing the other models with equal misclassification costs applied (models 1-3).

Table 10.3 shows that when equal misclassification costs are applied (models 1-3), a higher overall accuracy model (based on the validation set) is produced using predicted solubility (from solubility model 1 in Table 10.2) as a molecular descriptor to predict permeability class. Although model 3 has a lower overall accuracy, its specificity is much higher and this could be due to the influence of the solubility model included in the permeability model (solubility model 2). In other words,

improving the prediction of poorly soluble compounds resulted in higher predictive accuracy for poorly permeable compounds according to Table 10.3.

When higher misclassification costs are applied to false positives in the permeability models, models 5 and 6 have better overall accuracy (SE X SP) for the validation set and the lowest CNMI for the training set was obtained by model 5. Overall, the application of higher misclassification costs to reduce false positives resulted in the increased specificity and lower misclassification errors (CNMI), but overall accuracy is lower in models 4-6 in comparison with models 1-3. As expected, model 6, which included predicted solubility from model 2 in Table 10.2, had a higher specificity due to the higher misclassification costs originally applied to the solubility model – which have been utilised to improve predictive accuracy for poorly permeable compounds.

10.3.2 Interpretation of Selected Solubility and Permeability Models

Solubility classification models were developed using the top 20 molecular descriptors. In addition, permeability models were developed using either the top 20 molecular descriptors (selected using random forest) or the top 20 molecular descriptors plus predicted solubility from solubility models built in this work. It must be noted that although the top 20 molecular descriptors were given as input to the algorithm that builds the C&RT tree, not all the molecular descriptors were used to build the decision trees, since the C&RT also performs an additional 'embedded' feature selection process, adding to the tree only attributes deemed relevant for class prediction by the algorithm (Newby *et al*., 2013b). Furthermore, some molecular descriptors can be used more than once in a C&RT tree, as discussed below. Figure 10.1 is the selected solubility model 1 based on the classification decision tree.

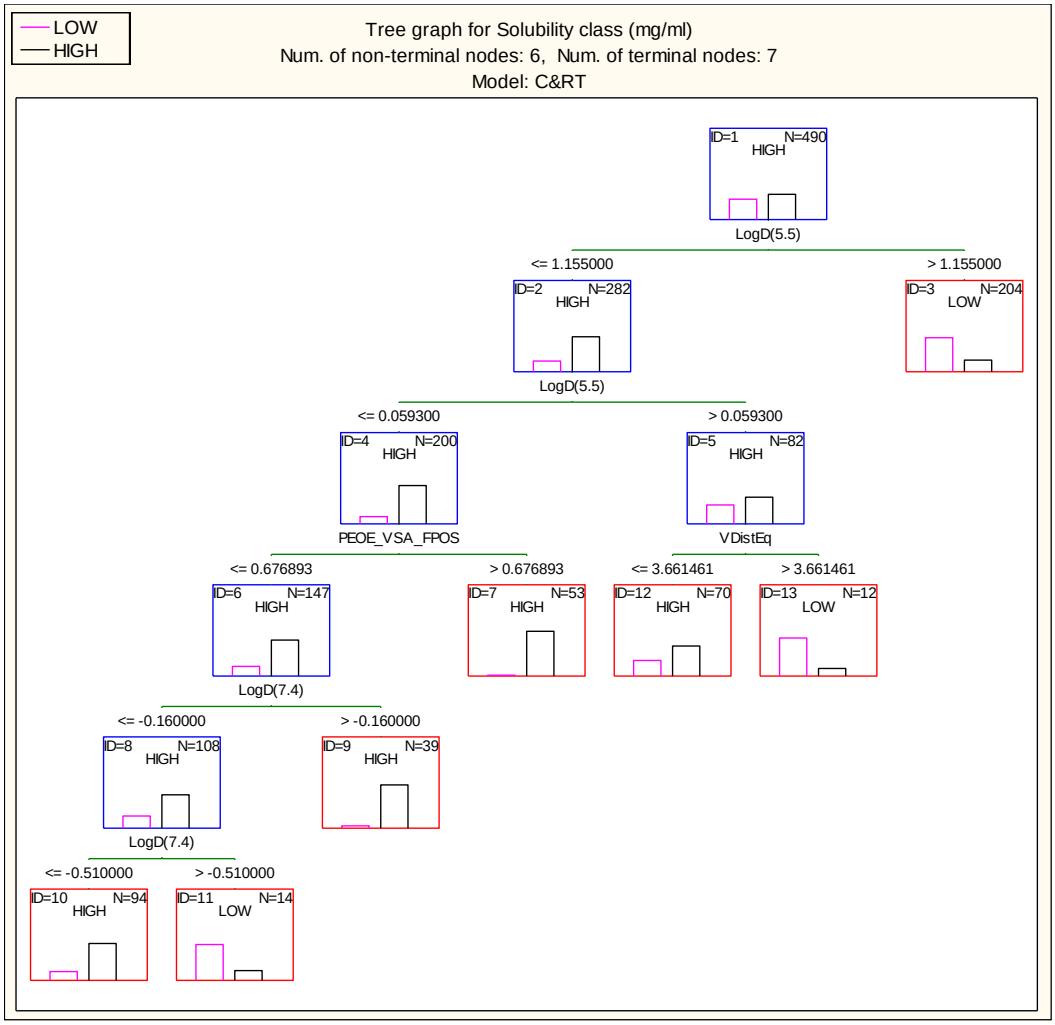

Figure 10.1. Tree graph for C&RT analysis for the prediction of solubility class with equal misclassification costs (model 1 in Table 10.2)

The first split variable in Figure 10.1 is ACDLogD(5.5), the logarithm of the apparent distribution coefficient between octanol and water at pH 5.5, a measure of hydrophobicity. This descriptor as well as logP has been used in many publications for modelling of different properties such as oral absorption (Ghafourian *et al*., 2012, Newby *et al*., 2013a), permeability (Gozalbes *et al*., 2011, Pham-The *et al*., 2013b) as well as solubility models (Gozalbes and Pineda-Lucena, 2010, Duchowicz *et al*., 2008). The use of logD at pH 5.5, despite solubility being measured at pH 7.4, is justified based on the fact that this descriptor indicates not only the effect of lipophilicity, but also the effect of acid/base property of the compounds. For example, an acidic and a basic compound of similar logP values will have different logD at this pH depending on their percentage of ionisation. At pH 5.5, the acidic

compound will be mainly unionised and hence its logD(5.5) will be close to its logP value, whereas the basic compound will be highly ionised therefore it will have a lower logD(5.5) than its logP value. In relation to solubility, highly lipophilic compounds can give rise to poor solubility, as indicated by Figure 10.1. In this model compounds are poorly soluble if they have a LogD(5.5) > 1.16 and examples of poorly soluble drugs in this node are diclofenac and ibuprofen – both are BCS class II compounds (poorly soluble but highly permeable) (Chuasuwan *et al*., 2009, Potthast *et al*., 2005). There is no further splitting of the highly lipophilic, poorly soluble compounds, indicating that this molecular descriptor is useful to define poor solubility (< 0.2 mg/mL) in this tree. The less lipophilic compounds (LogD(5.5) ≤ 1.16) are further characterised into high/low solubility using LogD(5.5); this time a lower threshold of 0.06 is used. In this case both nodes 4 and 5 are associated with high solubility; however, compounds that have higher LogD(5.5) (but lower than 1.16) are poorly soluble only if they have a vertex distance equality index (VDistEq) > 3.66. Computed from a distance matrix, VDistEq is mainly related to the size and shape (branching) of a molecule (MOE, 2014). Compounds with larger VDistEq tend to be larger and in most cases (less branched) linear molecules.

For compounds with lower LogD(5.5) than 0.06, the next molecular descriptor to split the tree is the partial charge descriptor, PEOE_VSA_FPOS. Using PEOE partial charge calculation (Gasteiger and Marsili, 1980), PEOE_VSA_FPOS is the sum of the van der Waals surface area of positively charged atoms divided by the total surface area of the molecule (MOE, 2014). According to Figure 10.1, those compounds with a PEOE_VSA_FPOS > 0.67 will be highly soluble, indicating that those with more positive partial charges (an indication of higher polarity and ionization) will be highly soluble. This is in agreement with the literature, where more polar molecules tend to be more soluble in water (Ghafourian and Bozorgi, 2010).

However, as depicted by this tree, node 6 (containing less polar compounds with PEOE_VSA_FPOS ≤ 0.67) is not pure at all and needs more splitting with other molecular descriptors; in this case, LogD(7.4) is used twice in the tree for these compounds. In Figure 10.1, compounds will be classed as poorly water soluble if

-0.51 < LogD(7.4) ≤ -0.16. It must be noted here that all these compounds have a LogD(5.5) below 1.155, as a result of division of node 2 and therefore they are hydrophilic enough to be classed as water soluble. Examples of these poorly water soluble compounds in node 14 are rofecoxib (Davies *et al*., 2003) and pindolol (Gazpio *et al*., 2005). Overall, from the solubility model, the main molecular descriptors used to classify solubility are those related to lipophilicity, ionization, polarity, size and shape, which is in accordance with the literature (Bergstrom *et al*., 2004, Bergstrom *et al*., 2002, Ghafourian and Bozorgi, 2010).

The best permeability model selected was model 2 in Table 10.3. Due to the size of the tree, in order to facilitate its interpretation the tree has been split into two trees (Figures 10.2 and 10.3). Figure 10.2 shows the half of the permeability decision tree that is built for those compounds predicted as poorly soluble by the solubility model 1 in Table 10.2. Figure 10.2 shows half of the C&RT tree for permeability built for those compounds predicted as highly soluble from the same solubility model. It must be noted that the trees in Figures 10.2 and 10.3 were originally one tree and the combined version, as well as all the other C&RT models presented in this work, is on the accompanying disk in this thesis.

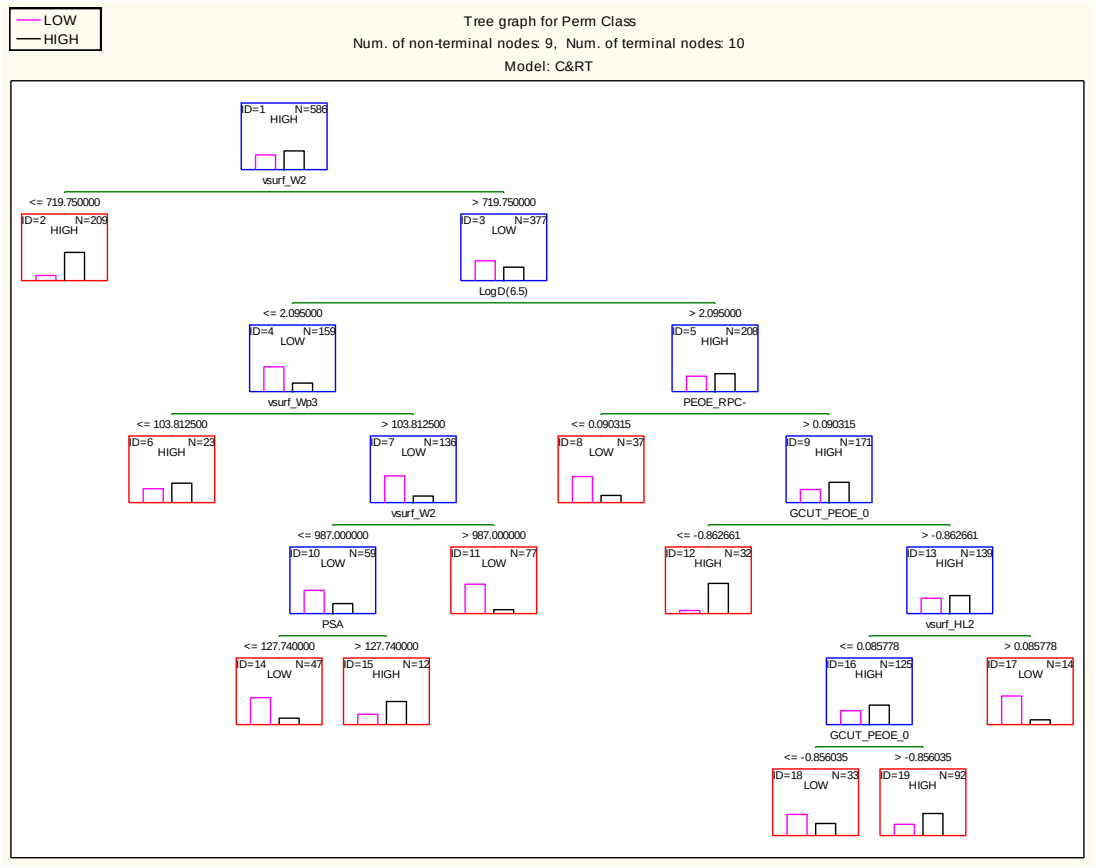

Figure 10.2. Tree graph for C&RT analysis (part of model 2 in Table 10.3) for the prediction of permeability class for predicted poorly soluble compounds from solubility model 1 (shown in Figure 10.1)

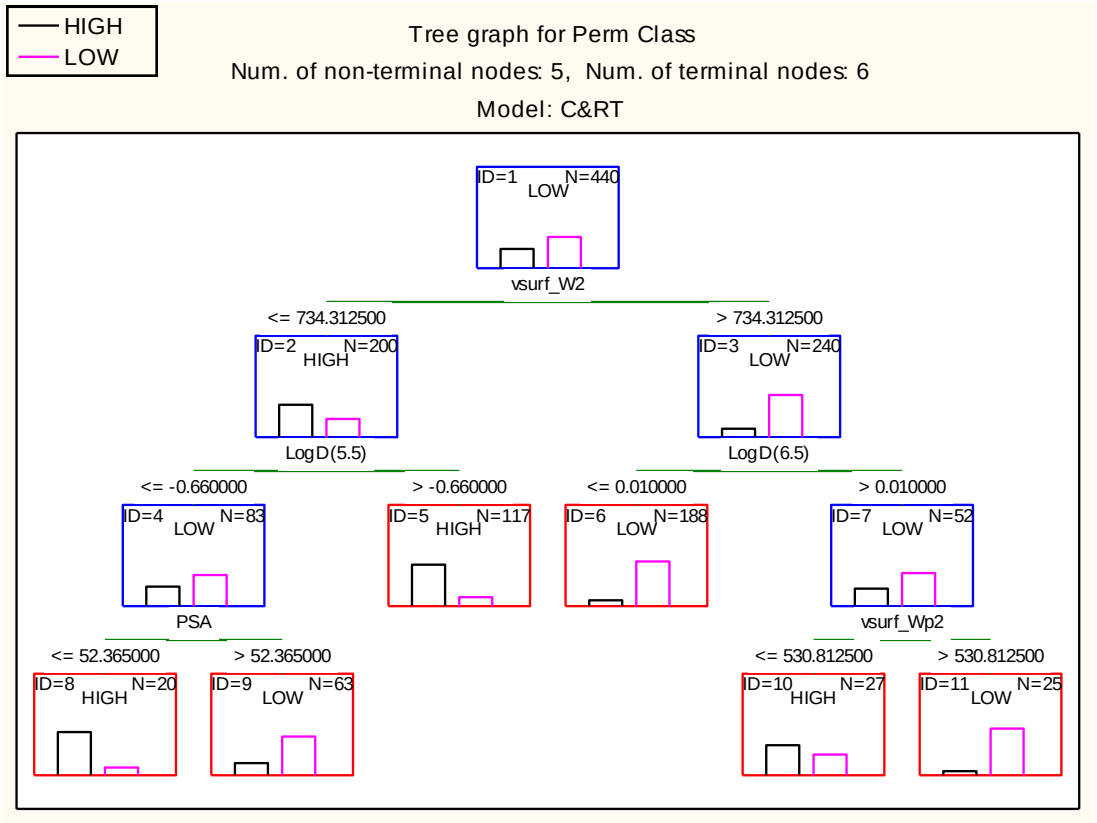

Figure 10.3 Tree graph for C&RT analysis (part of model 2 in Table 10.3) for the prediction of permeability class with equal misclassification costs for predicted highly soluble compounds from solubility model 1 (show in Figure 10.1)

Comparing Figures 10.2 and 10.3, it is noted that there is a slightly larger number of poorly soluble compounds (Figure 10.2) than highly water soluble compounds (Figure 10.3) in the permeability dataset and those poorly soluble compounds are mainly highly permeable (Figure 10.2) and *vice versa*. The first split of the tree in Figure 10.2 is using the vsurf_W2 molecular descriptor as calculated by MOE (MOE, 2012). Vsurf and related molecular descriptors are Volsurf descriptors described by Crucciani *et al* (Cruciani *et al*., 2000), which describe the size, shape, polarity, hydrophobicity and the balance between these properties on molecules. More specifically, vsurf_W descriptors describe the volume of hydrophilic regions of a molecule, calculated at certain interaction energy levels. In this case vsurf_W2, calculated at energy level 0.5 kcal/mol, accounts for the polarizability and dispersion forces in the hydrophilic regions of the molecules (MOE, 2012). According to this tree, poorly soluble compounds in Figure 10.2 will be classified as highly permeable as long as they have small hydrophilic volume (node 2). Compounds with larger

hydrophilic volumes in nodes 3 have been divided further according to logD6.5. In this case the general trend is that less lipophilic compounds (logD6.5 ≤ 2.10) will be mostly poorly permeable (node 4), which matches previous observations in caco-2 and other *in vitro* permeability cell lines (Sherer *et al*., 2012, Newby *et al*., 2014b). For those less lipophilic compounds (logD6.5 ≤ 2.10), the descriptor vsurf_Wp3 is used to discriminate between compounds with small polar volume (Vsurf_Wp3 ≤ 103.8) which are highly permeable, and compounds with large polar volume of the molecule (node 7). Compounds will be classified as poorly permeable due to their large polar volume unless they have smaller volume (Vsurf_W2 ≤ 987), but a polar surface area (PSA) greater than 127.7 (node 13). PSA has been cited to have a negative effect on oral absorption and hence permeability; this was also observed in previous chapters using oral absorption dataset. However, this is not what is presented in Figure 10.2 for the permeability dataset. The maximum PSA in this list of compounds (159 Å) is still moderate in comparison with the rest of the dataset. On closer inspection, the vast majority of these highly permeable compounds contain a sulphonamide or thiazole group. The polarity measure of these sulphur-containing functional groups using PSA seems to not correlate with the expected reduced absorption of polar compounds. Examples of these highly permeable compounds with large PSA values are glipizide and two oxazolidinones, antimicrobial agents PNU-182945 and PNU-183981.

For highly lipophilic compounds (logD 6.5 > 2.1) the next descriptor used to discriminate between high and low permeability is the relative negative partial charge descriptor calculated by PEOE (RPC-). This molecular descriptor is calculated by dividing the smallest (most negative) charge by the sum of negative charges on the whole molecule. Therefore, a higher number of hydrogen bond acceptors such as oxygen atoms in the molecules leads to lower values of RPC-. In this instance, compounds with a lower relative negative partial charge (≤ 0.09) are poorly permeable. Compounds with a higher RPC- are mainly highly permeable, but can be split further by the molecular descriptor GCUT_ PEOE_0. GCUT descriptors are calculated from a modified graph distance matrix using atomic partial charges calculated from PEOE method (see MOE helpfile, 2012). A minority of compounds with a lower GCUT-PEOE_0 than -0.86 have been classed as highly-absorbed.

These are structurally large and complex molecules with many rings and branches, mostly belonging to nucleotide based antivirals. Due to similarity of these compounds to natural metabolites, it is likely that they may have the possibility of being transported by carrier proteins.

Compounds with a higher GCUT_ PEOE_0 are also classified as highly permeable unless they have a vsurf_HL2 > 0.086 or, despite a smaller Vsurf_HL2, have GCUT_PEOE_0 ≤ -0.856. Vsurf_HL2 describes the hydrophilic-lipophilic balance, which is the calculated ratio between the hydrophilic regions measured at 4 kcal/mol and the hydrophobic regions measured at 0.8 kcal/mol (MOE, 2012). According to the tree in Figure 10.2, compounds are predicted as poorly permeable if they have a higher ratio of hydrophilic to lipophilic effect, and examples include bromocriptine and lansoprazole.

Figure 10.3 is the permeability model for compounds predicted as highly soluble according to solubility model 1. In this figure, the same top molecular descriptor as in Figure 10.2 is selected to split the compounds into high/low permeability in node 1. Compounds with vsurf_W2 values greater than 734.2, i.e. larger hydrophilic volume, are more likely to be poorly permeable according to this tree. This is unless they have a higher lipophilicity (logD6.5 > 0.01) and lower polar volume, according to vsurf_Wp2 ≤ 530.8. On the other side of the tree, the majority of compounds with relatively small hydrophilic volume are highly permeable, unless they are relatively hydrophilic at pH5.5 (LogD(5.5) ≤ -0.66) and have a PSA higher than 52.4. In this instance, this PSA threshold is similar to the threshold of 60 Å used for recent permeability modelling of Caco-2 permeability (Pham-The *et al*., 2013b). Based on Figures 10.2 and 10.3, it is interesting to note that the hydrophilic volume of a molecule is a better measure of permeability than the most widely known parameter, partition coefficient. For instance, in Figure 10.3, node 2, it can be seen that a good fraction of compounds with LogD(5.5) lower than -0.66 are highly permeable given the polar surface area is not too large (≤ 52.3).

10.3.3 Provisional BCS Class Prediction in an External Dataset Using Solubility and Permeability Models and Multi-Label Methods

The permeability and solubility models created previously were used to predict the BCS of an external validation set of 127 compounds with known values for both properties collected from the literature (BCS validation set). Different combinations of permeability and solubility models were tried in order to see what effect this would have on the overall results. Table 10.4 shows the results from the different combinations of the permeability and solubility models presented in Tables 10.2 and 10.3. For example, in Table 10.4, model 1 is the combination of the solubility model 1 (Table 10.2) and permeability model 1 (Table 10.3).

Recall that the multi-label method, binary relevance (BR), involves the prediction of permeability and solubility separately (models 1-2, 7-8 in Table 10.4), therefore it fails to take into account the relationship between these interrelated properties. Whereas the classifier chain (CC) method, which uses predicted solubility alongside structural molecular descriptors to help predict permeability, takes into account the label interactions (Models 3-6, 9-12 in Table 10.4). In Table 10.4, the overall accuracy (SP X SE) of the permeability and solubility models for the external validation set has also been included. In addition, the overall accuracy and geometric mean have been calculated alongside the individual class accuracies in order to help with interpretation.

Table 10.4. Results of the provisional BCS classification of an external validation set (n=127) to compare the binary relevance and classifier chain multi-label methods

Model	Multi-label method	Permeability Model Used (Table 10.3)	Solubility model Used (Table 10.2)	Permeability Accuracy (SP X SE)	Solubility Accuracy (SP X SE)	Overall Accuracy[a]	Geometric mean[b]	Class 1 accuracy (n=53) [c]	Class 2 accuracy (n=40) [c]	Class 3 accuracy (n=26) [c]	Class 4 accuracy (n=8)[c]
1	BR[d]	1	1	0.525	0.565	0.606	0.000	0.566	**0.725**	0.692	0.000
2			2		0.551	0.591	0.496	0.509	**0.725**	0.653	0.250
3	CC[e]	2	1	0.641	0.565	**0.630**	0.523	0.585	0.700	0.731	0.250
4			2		0.551	0.606	**0.590**	0.528	0.700	0.654	0.500
5	CC[e]	3	1	0.642	0.565	0.598	0.508	0.528	0.625	**0.806**	0.250
6			2		0.551	0.575	0.574	0.453	0.625	0.769	0.500
7	BR[d]	4	1	0.480	0.565	0.543	0.000	0.453	0.675	0.692	0.000
8			2		0.551	0.528	0.456	0.415	0.675	0.615	0.250
9	CC[e]	5	1	0.581	0.565	0.559	0.472	**0.604**	0.450	0.731	0.250
10			2		0.551	0.543	0.563	0.547	0.450	0.654	**0.625**
11	CC[e]	6	1	0.587	0.565	0.559	0.481	0.528	0.500	**0.808**	0.250
12			2		0.551	0.528	0.537	0.434	0.500	0.500	0.500

[a]Overall accuracy, calculated as correct number of predictions divided by total number of predictions; [b]Geometric mean, multiplication of all four accuracy measures of classes 1-4 and taking the fourth root of this product; [c]Class accuracy, number of compounds of a specific class that were correctly classified divided by total number of compounds in that specific class; [d]BR, Binary relevance; [e]CC, Classifier chain

From Table 10.4, based on the overall accuracy, i.e. the highest percentage of correct predictions, the best model to choose would be model 3. This model had an overall accuracy 0.630 (80/127) and was created combining the solubility model 1 and permeability model 2 (with incorporated predicted solubility). Although this model has the highest number of correct predictions, it has a poorer predictive accuracy for class 4. Therefore, using the geometric mean of the accuracy of all four classes, the best model would be model 4. This model was created combining the solubility model 2 and permeability model 2 (with incorporated predicted solubility). The difference between models 3 and 4 in Table 10.4 is the solubility model used with permeability model 2 to put compounds into BCS classes. Solubility model 1 from Table 10.2 is with equal misclassification costs and solubility model 2 is with higher misclassification costs to reduce false positives. Different combinations of the permeability and solubility models result in the different models having the best accuracy for all four classes. It is difficult to pick the best model based on the individual accuracies of the four classes. However, for overall accuracy the best model to choose would be either model 3 or model 4.

Models 1-6 were all derived from permeability models using equal misclassification costs applied, whereas Models 7-12 were derived from permeability models with higher misclassification costs applied to reduce false positives. Overall the application of higher misclassification costs to false positives in the permeability models (models 7-12) has led to lower overall accuracy and geometric mean accuracy; however, it has also led to the highest class accuracy for class 3 (model 11) and class 4 (model 10), due to better prediction of the low permeability compounds as expected.

In order to compare the models built by the two multi-label methods, firstly models 1 and 2 in Table 10.4 can be compared with models 3-6. Models 1 and 2 were built by the binary relevance method, whereas models 3-6 were built by the classifier chain multi-label method. Overall, based on the geometric mean, the classifier chain method obtained higher predictive ability across all classes. The only exception is that although models 5 and 6 have a higher geometric mean, they have a slightly lower overall accuracy compared with the binary relevance models 1 and 2. The superiority of the classifier chain method can also be seen from the permeability accuracy which was higher for the models built by the classifier chain method,

indicating that incorporating predicted solubility into models results in higher predictive accuracy for permeability. These patterns are also seen when comparing models 7-12, where higher misclassification costs have been applied to reduce false positives for the permeability models.

10.4 Discussion

This chapter has explored attempts to build permeability and solubility models to computationally predict a provisional BCS for chemicals in drug discovery by comparing two multi-label classification methods. The predictions can be very useful in early drug development and can streamline formulation and chemical optimization strategies. In addition, the BCS predictions can give insight into the mechanistic absorption properties of drugs, such as rate limiting steps like transporter effects or dissolution limiting solubility.

This work has involved multi-label classification of *in vitro* permeability and aqueous solubility to provisionally predict BCS classes for new chemical entities (NCEs) for early stage drug discovery. In order to compare the two multi-label methods, individual permeability and solubility models were built and validated. Initially, permeability and solubility models were built using the top 20 molecular descriptors as selected via random forest-based feature selection. Our previous study shows improved predictive accuracy when a pre-processing feature selection is performed prior to C&RT analysis (Newby *et al*., 2013b). In addition, permeability models were also built utilising the predicted solubility alongside the selected molecular descriptors to predict permeability class. The use of higher misclassification costs for false positives was also investigated to help improve class prediction of the poorly permeable and poorly soluble classes. Using the BCS validation set with known solubility and *in vitro* permeability, the predictions of the permeability and solubility models were combined and compared to the observed experimental BCS class. In this way, I compared two multi-label methods using the BCS validation set. Binary relevance involves the combination of separate, independently-built solubility and permeability models; however this does not take into account the interactions between these two labels. In order to overcome this, I compared this method to the multi-label method classifier chain. This method, in

relation to this work, involved the incorporation of predicted solubility to build and predict permeability class, and in doing so this method takes into account the relationship between these properties. Therefore, I am exploring the idea that the classifier chain method can help improve permeability class prediction and in turn provisional BCS class prediction.

10.4.1 Individual Permeability and Solubility Models

Both permeability and solubility are important properties in drug discovery. However, both these properties individually are complex and can be difficult to model. Lack of high quality datasets for drug-like compounds can contribute to the difficulty in predictions. BCS Class prediction can overcome variable permeability and solubility data by predicting compounds' classes rather than specific values as a first initial drug screen. However, suitable thresholds for discriminating between high and low permeability/solubility must be selected.

Permeability is the rate of drug absorption through Caco-2 cell line and is highly correlated with intestinal absorption (Newby *et al*., 2014b). Similar to intestinal absorption, there are many factors affecting and influencing permeability. According to the results of this study using the top 20 molecular descriptors from feature selection, permeability classes can be predicted with good accuracy. On the whole it is easier to predict the high permeability class than it is to predict the poor permeability class when equal misclassification costs were applied on a dataset with balanced class distribution (higher sensitivity than specificity values in Table 10.3). The same pattern emerges in relation to solubility, where according to this work better predictive accuracy is obtained for highly soluble compounds when using equal misclassification costs (Table 10.2). Solubility is also another complex parameter to predict with many complex interlinking factors (Salahinejad *et al*., 2013, Ghafourian and Bozorgi, 2010).

When equal misclassification costs have been applied, using predicted solubility as a molecular descriptor alongside the other molecular descriptors to build permeability models caused two things: models had better overall accuracy and better accuracy for poorly permeable compounds, in comparison with the model not incorporating

predicted solubility (see Table 10.3). Therefore, the inclusion of predicted solubility in this way increased the predictive accuracy of the poor permeability class. When higher misclassification costs were applied to improve the prediction of poorly permeable compounds, the specificity of permeability models also increased upon incorporating predicted solubility. Therefore, inclusion of predicted solubility into permeability models has resulted in better models or those that can predict poor permeability class better. This follows on from previous research whereby incorporating experimental permeability and experimental and predicted solubility into oral absorption models results in higher predictive accuracy (Newby *et al*., 2014b). When higher misclassification costs were applied to reduce false positives for the permeability models, overall lower predictive accuracy was observed. This could be due to the balanced nature of the dataset, containing roughly 50:50 high:low permeability compounds.

In terms of chemical space, for the solubility and permeability datasets, there seems to be a good overlap between the molecular properties (i.e. the top 20 molecular descriptors) for the training sets, individual validation sets and the BCS validation set for both these labels. This is indicated by a visual inspection of the scores plots from the principal component analysis which can be found in Appendix 4 (Figures A4.1 to A4.3). In addition, the applicability domain of the datasets was defined based on the Euclidean distances nodes, using KNIME v 2.9.4 (KNIME.COM AG), and showed that for solubility, only one compound was outside the applicability domain and none for the permeability dataset. However, more importantly, in this chapter, the solubility models were used to predict solubility for the permeability set. Therefore, it was important to check that the chemical space for solubility prediction of the permeability dataset was within the applicability domain. For the permeability dataset, only 20 compounds were outside the applicability domain of the solubility training set. A list of these compounds can be found in Appendix 4 (Table A4.4). Out of these 20 compounds, although outside the applicability domain for solubility prediction, 80% of these compounds were still classified correctly for permeability class using the best binary relevance and classifier chain models from Table 10.4.

10.4.2 Comparison of Most Relevant Molecular Descriptors

It is difficult to directly compare different permeability and solubility models used in the literature; however the molecular descriptor subsets used in the models can be compared. The top 20 molecular descriptors selected by random forest using predictor importance can be found in the Appendix 4 (Table A4.1 and Table A.4.2). In addition, the top descriptors chosen from the pool of 20, by the C&RT analysis, for the two properties can also be compared to see if there are similarities and/or differences, and this can be related back to the property in question. The top molecular descriptors selected by the solubility and permeability (C&RT) models are shown in Tables 10.5 and 10.6, respectively. The top molecular descriptors are counted by how many models they appear in. Also noted in Table 10.5 is if the molecular descriptor occurs more than once in the same decision tree. For Table 10.5, the molecular descriptors from solubility models 1 and 2 (Table 10.2) were used to show the top solubility molecular descriptors. For Table 10.6, permeability models 1 and 4 and models 2, 3, 5 and 6 (Table 10.3) were used to show the top molecular descriptors for the binary relevance and classifier chain methods respectively.

Table 10.5. The top molecular descriptors selected by C&RT for the prediction of solubility class (models 1 and 2 in Table 10.2)

Type of descriptor	**Descriptor**	**Number of C&RT models**	**Model (From Table 10.2)**
Lipophilicity	LogD(5.5)	4[a]	1,2
	LogD(7.4)	3[a]	1,2
Size/shape	VDistEq	3[a]	1,2
	BCUT_PEOE_0	1	2
	BCUT_SLOGP_2	1	2
Polarity/ Polarization	PEOE_VSA_FPOS	1	1
	PEOE_VSA_POL	1	2
Hydrogen bonding	MaxHp	1	2

[a]Occurred more than once in a single tree model.

For the solubility models, the top molecular descriptors (Table 10.5) picked by C&RT analysis was LogD(5.5). Other studies have identified lipophilicity

descriptors relating to LogD(5.5) and LogD(7.4), such as logP, as important for the prediction of solubility (Duchowicz *et al*., 2008, Jain and Yalkowsky, 2001). The next most frequently picked molecular descriptor is VDistEq, relating to the size and shape of the molecule. Larger molecules in drugs and drug like molecules tend to have higher lipophilicity (Ghafourian and Bozorgi, 2010) and additionally require higher energy to create a cavity in the solvent and solvate (solvation limiting solubility) (Wassvik *et al*., 2008). Additionally, the size and shape of a molecule can result in a rigidity that can cause high crystal lattice energy resulting in poor solubility (solid-state limiting solubility) (Wassvik *et al*., 2008, Ghafourian and Bozorgi, 2010). Finally, those descriptors relating to polarity and hydrogen bonding are also important for solubility prediction (Ghafourian and Bozorgi, 2010, Nelson and Jurs, 1994). Overall, molecular descriptors relating to lipophilicity, size, shape, polarity and hydrogen bonding are all important for solubility of drug compounds as they relate to the crystal lattice energy, solvent cavity formation energy and solvation energy – all important factors for solubility of drug compounds (Ghafourian and Bozorgi, 2010, Nelson and Jurs, 1994, Hewitt *et al*., 2009).

Table 10.6. The top molecular descriptors selected by C&RT for the prediction of permeability class for the binary relevance (models 1 and 4, Table 10.3) and classifier chain permeability models (models 2, 3, 5 and 6, Table 10.3)

Type of descriptor	Descriptor	BR permeability models		CC permeability models	
		Number of C&RT models	Model (From Table 10.3)	Number of C&RT models	Model (From Table 10.3)
Lipophilicity/ Hydrophobicity	LogD(6.5)	3[a]	1,4	6[a]	2,3,5,6
	LogD(5.5)	1	1	3	2,3,6
	LogD(10)			3	3,5,6
	LogD(7.4)	2	1,4		
	vsurf_HL1	2	1,4		
	vsurf_HL2			4	2,3,5,6
	vsurf_CW4	1	1		
Size of hydrophilic/polar regions	vsurf_Wp3	2	1,4	7[a]	2,3,6
	vsurf_W2	1	1	7[a]	2,3,5,6
	vsurf_W3	1	4	2	5,6
	vsurf_Wp2	1	4	2	2,5
	PEOE_RPC-	1	4	4	2,3,5,6
	PSA			2[a]	2
Size/Shape	xv2	2[a]	4		
	GCUT_PEOE_0	3[a]	1,4	8[a]	2,3,5,6
	chi1_C	2	1,4		
Bascity	FIBpH6.5	2[a]	4		
Hydrogen bonding	vsurf_HB1	5[a]	1,4	3[a]	5

[a]Occurred more than once in a single tree model.

The top molecular descriptors for the permeability models in this work picked by the resulting C&RT analysis can be roughly grouped into five groups: lipophilicity/hydrophobicity parameters, those describing the size of the hydrophilic or polar molecular regions, basicity, hydrogen bonding, and finally size/shape parameters (Table 10.6). Overall, there are 25 cases of lipophilicity/hydrophobicity parameters used in the permeability models and 30 cases of parameters describing the size of the hydrophilic or polar regions of the molecule. These two make up 69%

of permeability related features. There are only two instances of the basicity parameter, eight cases of hydrogen bond donor effect and 15 cases of molecular descriptors relating to size and/or shape utilized in the permeability models. The importance of hydrophilic or polar size of the molecule has been seen in previous literature. In particular, polar surface area has been cited to be important for permeability classification between low, medium and high permeability, and is a popular molecular descriptor used in my models (Pham-The *et al*., 2013b). Molecular descriptors relating to hydrogen bonding are also popular in relation to permeability (Nordqvist *et al*., 2004) as well as oral absorption. More specifically, hydrogen bonding is two of the descriptors used in the widely accepted filter for identifying poorly-absorbed compounds, Lipinski's rule of five (Lipinski *et al*., 1997). Molecular descriptors which are important for permeability such as those relating to lipophilicity, size/shape, polarity and hydrogen bonding are also important for the prediction of oral absorption (Newby *et al*., 2013a, Newby *et al*., 2013b, Ghafourian *et al*., 2012).

10.4.3 Comparison with Related Literature

To my knowledge there are few studies which use QSAR models to predict BCS class. However, there are many individual studies that predict either permeability or solubility. A related work has been published recently by Pham-The *et al*. (2013), which is different from this study in terms of the methods, parameters used and property thresholds.

As a solubility measure, Pham-The *et al*. used dose number (D_o), defined as the ratio of drug concentration following a given dose in the stomach of 250ml volume to the saturated solubility. One of the problems with using D_o for a provisional prediction is that D_o is a property of the drug formulation and not a specific property of the active compound. Therefore, the maximum dose can depend on many things such as formulation type, toxicity and drug target affinity or even different doses of drug may be used to treat different disease severities or even different disease states (Broccatelli *et al*., 2012). In terms of future predictions, maximum dose will be needed from literature in order to calculate D_o. The advantage of my models

described here is that they do not need any experimental values such as the drug dose for future predictions.

They also used a permeability threshold of 16 x 10^{-6} cm/s, based on the permeability of metoprolol, a highly-absorbed drug. This threshold is over double the threshold that was objectively selected and statistically validated using the correlation between oral absorption and *in vitro* permeability in previous studies (Newby *et al*., 2014b). The individual permeability and solubility models developed by Pham-The *et al* using a dataset of 322 compounds achieved good overall accuracy for the training and validation sets (>75%). Due to the different datasets and validation and training sets, the accuracy of the models cannot be directly compared. I have used larger datasets for model development that cover a large chemical space. In addition, the different thresholds used lead to different classification problems, each resulting in different levels of difficulty for classification of each property.

Pham-The *et al*. (2013) validated the models by using firstly an external validation set containing 57 compounds from the WHO (World Health Organization) list of essential medicines. Unfortunately, in this validation set there were no experimental Caco-2 permeability data to validate the permeability prediction. Furthermore, over half of these compounds are assigned into more than one class, which is potentially inconclusive. My work involved an validation set to validate permeability and solubility models and in addition an BCS validation set where both permeability and solubility were known, in order to validate BCS prediction.

There are studies in the literature that predict BDDCS class (Biopharmaceutics Drug Disposition Classification System) (Wu and Benet, 2005) instead of BCS class. The BDDCS system classifies compounds into one of the four BDDCS classes based on the rate of metabolism, instead of permeability used in the BCS system, and solubility (using dose number). There appears to be a correlation between BCS and BDDCS classes, but only for passively absorbed compounds (Broccatelli *et al*., 2012). With the growing number of compounds being identified as undergoing carrier mediated absorption, the comparison of BCS and BDDCS models could be complicated.

10.4.4 Comparison of BCS Class Assignments with the Literature

The external validation set of 127 compounds contained both *in vitro* permeability and aqueous solubility collected from the literature. Based on the literature data, an observed BCS class was assigned to these compounds using my thresholds for permeability and solubility. Searching the literature, I found reported BCS classes for 71 of the 127 compounds in the validation set. From these 71, 10 compounds were cited in the literature to belong to more than one class and 16 were cited to belong to a different class from what I had assigned them based on my solubility and permeability thresholds. Different assignments of BCS class to compounds in the literature have also been shown in other studies (Bergstrom *et al*., 2014).

On closer inspection of these 16 compounds, the main differences between my assigned BCS class and the literature-assigned BCS class are the effect of maximum dose and pH, which have not been considered in my work. In addition, there are *in vitro – in vivo* differences due to varying levels of transporter expression in cell lines and gastrointestinal tract. As a result, some compounds that are poorly soluble and poorly permeable or highly permeable but poorly soluble *in vitro* may not necessarily be poorly-absorbed *in vivo*. Examples include cinacalcet (Class IV), which is poorly soluble and poorly permeable but is absorbed >80% and dapsone (Class II) which is poorly soluble but has a %HIA of 90%. The external validation set with the experimentally (*in vitro*) assigned and literature assigned compounds can be found in Appendix 4.

Concerning the 10 compounds cited as belonging to more than one class, it is interesting to see how the best models (those with the best overall accuracy and geometric mean accuracy, i.e. models 3 and 4 in Table 10.4) predicted these compounds, as their prediction may give more evidence to the assignment of these compounds to that class. For example based on my experimental data, ethosuximide is classified as belonging to class I, however the WHO guidelines state that the classification of this compound could be either class I or class III due to insufficient data on permeability. The models 3 and 4 from Table 10.4 both predict that this compound is class I, and this is supported by a %HIA of 93%. For the rest of the

compounds, the majority are predicted into either one of the cited classes by models 3 and 4.

Using model 4 from Table 10.4, it is interesting to see which class was assigned to the compounds in BCS validation set. This can help understand the error rates associated with the model and the tendency of the model in relation to BCS class prediction. This confusion matrix comparing predicted versus observed BCS classes is shown in Table 10.7.

Table 10.7. Confusion matrix of model 4 from Table 10.4 for the prediction of BCS classes for the validation set

	Predicted Class 1	**Predicted Class 2**	**Predicted Class 3**	**Predicted Class 4**	**Total**	**Accuracy (%)**
Observed class 1	**28**	15**	6**	4**	53	52.8
Observed class 2	7*	**28**	1	4**	40	70.0
Observed class 3	4*	1	**17**	4**	26	65.4
Observed class 4	1*	2*	1*	**4**	8	50.0
Total Compounds	40	46	25	16		
Precision (%)	70.0	60.9	68.0	25.0		

*Type I errors
**Type II errors
Precision (%) is calculated for each class by adding the number of compounds in the column for that class and dividing by the total number of compounds (column total) for that class. Accuracy (%) is calculated by adding the number of compounds for each class in the row for that class and divided by the total number of compounds (row total) for that class.

Type I and type II errors were calculated for the values reported in Table 10.7. According to Khandelwal *et al*. (Khandelwal *et al*., 2007), Type I errors (false positive errors) represent those compounds that are either predicted class I when in fact they are observed to be BCS classes II-IV or predicted class II or III but are actually class IV compounds. Therefore, the predicted class is biopharmaceutically more favourable than the observed actual class. Type II errors (false negative errors) represent those compounds that are either predicted as class IV but were observed to be BCS classes I-III, or are predicted as class II or III but were observed to be class I.

In other words, the predicted class is biopharmaceutically less favourable than the true class. The % of type I errors was 11.8 % and the % of type II errors was 25.9%. The results from a similar study by Pharm-The *et al*. (Pham-The *et al*., 2013a) calculated type I and type errors II of 10.6% and 14.6% respectively, for their entire dataset (training and validation set) of 322 compounds.

It has been proposed that for BCS class prediction type II errors should be kept as low as possible (Pham-The *et al*., 2013a). This is quite obvious given that BCS class is used for the decision making regarding biopharmaceuticals experimentations required for oral dosage forms. Additionally, it might be more desirable to have good precision of class I compounds, rather than good accuracy, as these compounds are prioritised for biowaivers (CDER/FDA, 2000). This principle of focussing on precision rather than accuracy may be appropriate for class III compounds too, due to the increasing evidence for the suitability of class III compounds for biowaivers (Crison *et al*., 2012). As seen in Table 10.7 both of the precision measures for class I and III were higher than the respective accuracy measures. Based on this, it is interesting to see that, although class III is not the most popular represented class in the external validation set compared with classes I and II, it still has high class accuracy and precision.

It is important to state that the main difficulty for the models in this work was encountered in predicting class IV compounds. This was not entirely unexpected, since although the permeability and solubility datasets had balanced class distributions, the combination of these resulted in an under-representation of class IV. This may not be a major concern for industry; however, from a prediction point of view, by not considering the predictive accuracy of all classes can result in a higher number of misclassifications, which could prove costly for industry (Khandelwal *et al*., 2007). This could be resolved by balancing all four BCS classes; however this can drastically reduce the number of compounds and potentially the models' ability to predict new compounds. My work has utilised all data available and applied misclassification costs to attempt to overcome the BCS class imbalance. However, the poor prediction may not be down to the poor representation of classes

and could be also a result of self-association in water, as cited in other research (Broccatelli *et al*., 2012, Ross and Riley, 1990).

10.5 Conclusion

The *in silico* prediction of a provisional BCS class is a challenging task. One of the challenging aspects of BCS class predictions is the potential effect of solubility on permeability prediction. Separate models of permeability and solubility fail to take into account the interactions between the class labels, and modelling each label separately reduces the generalisation for new compounds. It is well known in the literature that poor solubility can give rise to poor and variable absorption. Therefore, permeability prediction should include and so take into account the effects of solubility. Hence, using predicted solubility in permeability models alongside structural molecular descriptors, as performed in this work using the classifier chain multi-label classification method, avoids the disadvantage of other modelling methods for BCS prediction, like binary relevance multi-label classification.

This work has shown that the classifier chain multi-label method can greatly influence permeability models and hence provisional BCS using C&RT analysis. The use of predicted solubility as a descriptor to build and predict permeability, using the classifier chain method, has been shown to improve a permeability model's predictive accuracy and in turn final provisional BCS prediction. The molecular descriptors used by both solubility and permeability models relate to lipophilicity, hydrogen bonding, polarity, size and shape; however their relationship with these properties is usually inversely related.

The benefit of the binary relevance and classifier chain methods over algorithm adaption methods is the utilisation of large datasets for permeability and solubility. There was no restriction to the dataset just because of missing values, as separate models for permeability and solubility were built based on the available data for each property. One limitation with this type of protocol is the lack of generalisation for the poorly represented class IV compounds. However, this can be improved slightly with the application of higher misclassification costs. The literature reveals a lack of multi-label classification methods for provisional prediction of BCS class suitable

for a drug discovery scenario. Therefore, according to my results, the classifier chain method can be used successfully to improve the prediction of permeability class using predicted solubility.

Future extensions to this work would be to utilise more types of multi-label classification methods to perform consensus prediction similar to those in the literature; however the method must be able to include and use predicted solubility in the permeability model.

In conclusion, this work has highlighted the potential benefit of using the classifier chain multi-label method, to predict provisional BCS class prediction for drug discovery (Newby *et al*., 2014a).

11 Conclusions

The chapter aims to give a summary of the overall conclusions of this research and how the author has contributed to existing knowledge in relation to the aims set at the beginning of this thesis.

With the increase in cost, there has never been a more demanding time in the pharmaceutical industry to increase cost effectiveness in drug discovery. *In silico* techniques offer inexpensive methods to assist drug discovery to fail fast and cheaply and identify promising compounds. This thesis has covered various aspects of Quantitative Structure-Activity Relationship (QSAR) including training set selection, model choice, feature selection, model optimisation and appropriate statistics for model validation. In this research, QSAR methods were employed to model oral absorption and related properties.

Initially, using a previously published dataset of oral absorption data, two methods were utilised to overcome the problem of unbalanced class distribution for oral absorption. The methods used to overcome this problem, namely under-sampling of the majority class (chapter 6) and higher misclassification costs for the minority class (chapter 7), have both been shown to overcome the unbalanced class distribution by producing models with higher accuracy and applicable to an industry scenario.

The impact of under-sampling the majority class in the training set to improve model accuracy has been shown to be successful using linear and non-linear techniques for numerical and categorical prediction of oral absorption. The use of MLR and LDA offered linear methods (chapter 6) and decision trees using C&RT offered a non-linear method (chapter 7) to indicate the successfulness of the under sampling technique using a variety of methods. Although the application of this method for oral absorption datasets obtained a positive result, the reduction of data using this technique is not ideal and could cause problems with generalization to new compound sets, especially for small datasets.

Therefore, the application of higher misclassification costs for minority class was investigated in chapter 7, in order to overcome the biased class distribution of the highly-absorbed compounds without data removal. When higher misclassification

costs were applied to reduce false positives, higher predictive accuracy models were produced, with higher accuracy for the underrepresented poor absorption minority class as shown in chapters 7 and 8. The results of these two chapters led to the application of higher misclassification to overcome unbalanced class distribution in chapter 9, to help elucidate the effect of solubility and permeability on oral absorption. Additionally, chapter 7 highlighted the use of higher misclassification costs on a dataset with a balanced class distribution. The application of higher misclassification costs for either class provides the opportunity to improve the predictive accuracy for one class when using a balanced dataset, depending on the purpose and motive of the resulting model. This principle was applied to permeability and solubility models in chapter 10, which both had balanced datasets. Due to the growing number of poorly-absorbed compounds with poor permeability and/or poor solubility on the market, there was an incentive to build models with higher misclassification costs for these properties even though the dataset was balanced. Therefore, application of misclassification costs not only offers a way to improve model accuracy by increasing the predictive ability of the minority class but also provides the user with the opportunity to apply costs depending on the incentives of oral absorption model prediction in drug discovery and development.

The ultimate goal of QSARs for oral absorption is to provide the most accurate prediction of this property by using the most appropriate methods for this particular dataset. One approach to achieve this is through optimization of the feature selection process to find the most significant molecular descriptors for the prediction of oral absorption. Molecular descriptors are the building blocks of QSAR models; hence the most important ones need to be selected. There are a large variety of molecular descriptors (or features) to choose from; however, the general assumption is that feature selection methods reduce the number of molecular descriptors to be used in a model and result in models with increased interpretability and predictive power. Various feature selection methods were utilised in this thesis. In chapter 6, the use of stepwise regression, in order to reduce the number of variables, for numerical prediction using MLR, led to some subsets of molecular descriptors, plus other subsets based on Lipinski's rule of five, being utilised in later models in chapter 7. The classification C&RT models built using these subsets of molecular descriptors in

chapter 7 overall had similar or higher predictive accuracy and increased interpretability due to the reduction of molecular descriptors, compared with models built when the C&RT algorithm was able to select from all molecular descriptors available. This conclusion was further investigated and confirmed in chapter 8, where five 'pre-processing' filter feature selection methods were utilised to select molecular descriptor subsets for oral absorption classification. These subsets of molecular descriptors were used to build classification models using C&RT which is an embedded feature selection approach in its own right. The resulting models were compared with models built with no pre-processing feature selection and relying on the embedded feature selection in C&RT. It is apparent from the results of chapters 7 and 8 that different feature selection methods utilised should be tried and evaluated for a given dataset. One important aspect of the feature selection work was the combination of feature selection alongside the application of higher misclassification costs as utilised in chapter 7. In addition the application of higher misclassification costs to one of the feature selection methods, namely variable importance using random forest, resulted in high predictive accuracy models that also coped well with the class imbalance problem discussed previously. The success of the variable importance using random forest gave the justification of using this feature selection approach to select molecular descriptor subsets for permeability and solubility prediction in chapter 10.

The remaining chapters focus on oral absorption related to the two fundamental properties that govern the rate and extent of oral absorption: permeability and solubility. The conclusions from previous chapters influenced the development of resulting models such as the application of higher misclassification costs in chapters 9 and 10 and the application of the pre-processing feature selection method, variable importance, utilised in chapter 10 to improve model accuracy based on the results of previous chapters.

The availability experimental data of permeability and solubility in drug discovery is growing rapidly, whereas the number of available human intestinal absorption data remains mostly unchanged. Based on chapter 6, it was realised that the impact of solubility would be an important factor to consider in relation to oral absorption prediction. Therefore, it was shown in chapter 9 and 10 that models that utilise high/

moderate throughput data generated in drug discovery can aid in the predictive accuracy and interpretation of oral absorption models by utilising the relationship these properties share with oral absorption. Permeability was shown to have good influence on the prediction of oral absorption. Furthermore, the inclusion of experimental permeability and solubility values increased the accuracy of oral absorption prediction even more. In spite of this, the interaction between oral absorption and solubility in chapter 9 was not as apparent as expected, despite the known close relationship between absorption and solubility.

The results from chapter 9 confirm previous conclusions in the literature where *in vitro* permeability has a good correlation with human oral absorption, despite data variation among different laboratories. The relationship between high permeability and high absorption was strongest and the relationship between poor absorption and poor permeability was less certain. This could be due to lack of poor absorption data available in the literature. The application of higher misclassification costs utilised in the previous chapters 7 and 8 was shown to improve the predictive ability of an objective permeability threshold to predict oral absorption class.

Using the objective permeability thresholds, oral absorption models were built using decision trees, utilising solubility and related parameters, as well as the calculated molecular descriptors to predict oral absorption class. Models that utilised predicted solubility using the general solubility equation (GSE) and the melting point based absorption potential (MPbAP) gave rise to better predictive models and could therefore potentially be used if experimental solubility was not available for the prediction of oral absorption class. The models developed in chapter 9 extend those published in the literature by incorporating all compounds including those with solubility issues. Although the lack of negative impact of poor solubility on oral absorption was unexpected, these models were able to predict the oral absorption of those compounds identified with solubility issues. Therefore the incorporation of solubility into oral absorption models alongside permeability can help classify oral absorption.

The relationship between absorption/permeability with solubility was investigated further by using multi-label classification for the prediction of BCS

(Biopharmaceutics Classification System) class in chapter 10. As the BCS assigns a class according to the permeability and solubility and these are considered the fundamental properties governing absorption, it seemed fitting to utilise the dataset collected in chapter 9 to model these two fundamental properties in a multi-label fashion. The multi-label methods introduced in chapter 10 are used to predict permeability and solubility models, which in turn will give a provisional BCS class prediction to help with formulation and chemical optimisation in drug discovery and development.

In chapter 10, the comparison of two multi-label methods was investigated. The first method was binary relevance, which did not take into account the interaction between solubility and permeability. The second method was classifier chain, which takes into account label interactions between solubility and permeability by including predicted solubility as a feature, to build the permeability models in this work. The classifier chain method was utilised based on the results of chapter 9, where the inclusion of solubility and related parameters help to predict oral absorption class. The influence of solubility was shown to impact and improve permeability prediction models from the work presented in chapter 10. In this chapter, the classifier chain method had higher predictive accuracy for the prediction of BCS using an external validation set of 127 compounds.

Additionally, the use of higher misclassification costs to reduce false positives in this chapter 10 yielded results which had lower overall accuracy for solubility and permeability models. However, all models with higher misclassification costs applied had better predictive accuracy for the poor permeability/solubility classes. In spite of the lower accuracy, the classifier chain method still resulted in higher geometric mean accuracy compared with the binary relevance method. The use of higher misclassification costs for models resulted in higher predictive accuracy for the underrepresented classes III and IV in the BCS external validation set. This chapter has highlighted the potential benefit of using the classifier chain multi-label method to predict provisional BCS class for drug discovery.

The molecular descriptors utilised in the models throughout this thesis, such as those describing size, shape, polarizability and hydrogen bonding, can be related to the

known mechanisms of permeability, solubility and oral absorption. The later models in chapters 9 and 10 are useful for a better mechanistic understanding of the effect of these properties as well as for identifying molecular descriptors that contribute to overall oral absorption.

Based on the current results, it can be concluded that modelling using QSAR can aid the understanding and interpretation of the mechanisms involved in the oral absorption of drug compounds and with the appropriate data can lead to validated models with good predictive accuracy. Furthermore, the *in silico* modelling approaches and models presented in this work can aid a variety of disciplines ranging from medicinal chemistry through to drug formulation. It is also expected that the modelling approaches used for the prediction of oral absorption could be applied for the prediction of other pharmacokinetic properties in the pharmaceutical industry in order to build interpretable predictive models analogous to the ones presented in this thesis.

12 Future Research Directions

The research presented in this thesis has raised many interesting questions. Therefore, there are several lines of research arising from this work which could be pursued further.

Firstly, although all the models presented in this thesis have been externally validated, it will be essential to re-validate the models with new compound data when available, additionally checking the applicability domain of those new compounds to make sure they are not extrapolated based on the chemical area of the training set. The lack of experimental data added to oral absorption, permeability and solubility datasets could be resolved by using semi supervised learning techniques (Chapelle *et al*., 2006, Li *et al*., 2008), as well as modelling the data already used in this thesis with approaches other than decision tree analysis such as neural networks and support vector machine. Furthermore, an ensemble of classifiers could be used in order to achieve higher predictive accuracy for the prediction of oral absorption based on a multitude of different QSAR methods.

Overcoming the biased class distributions in oral absorption datasets using the approaches presented in this thesis was shown to be successful. In addition to under-sampling the training set and the application of higher misclassification costs to the minority class, a technique called over-sampling could have been carried out to overcome the biased class distribution. Over-sampling is similar to the concept of under-sampling, but instead of compounds of the majority class being removed to create a balanced dataset, the minority class is expanded by duplication of these minority compounds to match the number of compounds in the majority class in the dataset. Although over-sampling creates a balanced class dataset, it does not add any information due to the replication of the minority class compounds. This could result in overfitting of models especially with decision tree methods (Barandela *et al*., 2004, Zang *et al*., 2013). In spite of this, this method could be an interesting comparison with the methods presented in this work, as it has been suggested that over-sampling has similar predictive accuracy to under-sampling and misclassification costs in some cases (Zang *et al*., 2013).

In this work there have been several feature selection approaches presented. There are still many other feature selection methods that could be explored. In chapter 8, filter feature selection methods were compared, however wrapper feature selection methods were not considered, so this could offer another angle for molecular descriptor selection and model development (Eklund *et al*., 2014). There are many types of methods that could be investigated such as those inspired by nature like particle swarm optimisation, ant colony optimisation and reshaped sequential replacement method, which has been introduced as a promising method adapted from the original sequential replacement method (Grisoni *et al*., 2014, Correa *et al*., 2008). Additionally, there has been work exploring feature selection methods for multi-label datasets, therefore this is a promising extension to the multi-label work carried out in chapter 10 (Jungjit *et al*., 2013, Spolaôr *et al*., 2013). Based on this research and the literature, different feature selection methods need to be compared for the same dataset (Xue *et al*., 2004). Moreover, there is scope for the feature selection approaches in this thesis to be applied to other pharmacokinetic datasets.

The impact of solubility on oral absorption model was investigated in chapter 9. Experimental aqueous solubility and predicted solubility and related values were used in order to help predict oral absorption. Although their impact was not as apparent as expected, it would have been interesting to use a number of other solubility and related parameters to confirm this observation. Firstly, solubility obtained from simulated intestinal and gastric fluid may offer a more real life scenario by considering pH and endogenous components, for oral absorption prediction. Secondly, the predicted solubility values used in this thesis, namely GSE and MPbAP, are both based on experimental melting point values. There are other solubility models presented in the literature that use calculated theoretical molecular descriptors (Raevsky *et al*., 2014). Predicted solubility model examples include ESOL (Delaney, 2004), SCRATCH (Jain and Yalkowsky, 2010) and GSE using TPSA instead of melting point (Ali *et al*., 2012). All of these examples are purely theoretical and not based on experimental data. Therefore, the calculation of predicted solubility from some of these equations would be useful for those compounds with no melting point available in the literature. However, it has been shown that the predicted solubility from the examples mentioned had lower model

accuracy compared with GSE which was used in this work. In spite of this, it would be potentially useful to see their ability as molecular descriptors for oral absorption.

Solubility is a complex property which can be described by many different properties. It is suggested that a better description of the solid-state properties will improve the predictive accuracy of the solubility models; therefore it would be interesting to investigate how these properties impact on oral absorption models. (Wassvik *et al*., 2006). A good example of solid state properties relating to solubility is melting point, as shown in chapter 9. Even though melting point was not shown to be useful on its own, there are potentially other solid state properties of solubility such as wettability and density that could be investigated as estimators of solubility, and the their impact on absorption prediction investigated.

Multi-label classification was introduced in chapter 10. Multi-label classification is a relatively new approach in the context of modelling pharmacokinetics properties using QSAR; therefore there is plenty of scope to expand this area further. Firstly this thesis presented two multi-label methods, binary relevance and classifier chain, however there are many more multi-label methods that could be used such as power set and algorithm-dependent methods (see below) (Carvalho and Freitas, 2009, Tsoumakas and Katakis, 2007). A problem with these methods is the unbalanced distribution of class labels, which results in poor accuracy due to under-representation of many labels. In spite of this, methods such as under sampling, oversampling and misclassification costs could be investigated for multi-label classification. This would be an interesting route as the power set method takes into account label interactions, and if the under-representation was overcome maybe this method could be highly predictive for BCS class prediction.

In addition, this thesis only considered problem transformation multi-label methods, i.e., the multi-label problem was converted into a single label problem and then modelling was carried out using decision trees. In the data mining literature, there are algorithm-dependent multi-label methods which develop models based on the multi-label data directly (Clare and King, 2001, Min-Ling and Zhi-Hua, 2005). Therefore, another aspect of future work could be the development and utilisation of algorithm-dependent methods for provisional prediction of BCS class. The

complexity of such a project would be advantageous for comparison with the multi label problem transformation approaches in a pharmaceutical sciences context. There is growing awareness of multi-label problems, therefore further investigations into multi-label classification will be required (Cherkasov et al., 2014).

Finally, there is great interest in carrier-mediated transporters, particularly those in the small intestine. Furthermore, the knowledge of substrate specificity for multiple transporters can be useful to help with compound design, chemical optimisation, and predicting drug-drug interactions which all can offer a mechanistic understanding of the role of intestinal transporters towards oral absorption. Therefore, data mining tools that can elucidate which drug compounds are substrates for intestinal transporters are highly desirable.

Carrier-mediated transporters can be categorised into pre-defined family classes which are naturally hierarchical. For example, transporters can firstly be classified into general types of transporter (e.g. ion channel, active transporter) then further classified into more specific gene sub families and so on until classified into a very specific individual transporter. Hierarchical classification approaches could be used, which would involve building classification models that can predict if drug compounds are absorbed via transporters, and if so, also predict the classes of its transporter, ranging from generic to very specific transporter classes in a pre-defined transporter class hierarchy as described previously. Hierarchical classification approaches take into account the similarities and hierarchical relationships among transporter classes, exploiting that information to improve predictive accuracy of models produced. It has been shown in the data mining area that hierarchical classification approaches in general have higher predictive accuracy compared with the conventional "flat" classification approaches used in this thesis, as well as offering more information regarding the hierarchical classes (Silla and Freitas, 2011, Cerri *et al*., 2011). Using the hierarchical classification approach could lead to development and optimisation of classifications models using a variety of feature selection and classification algorithms (Secker *et al*., 2010, Costa *et al*., 2007).

It is clear from this section that as the data mining community develops and produces new techniques, it will not be long before the QSAR community will utilise and

adapt them for their own needs. The future of QSAR will involve multidisciplinary research, resulting in models with better predictive accuracy and interpretability that will hopefully provide a better mechanistic understanding of oral absorption of drug compounds.

13 References

ABRAHAM, M. H., ZHAO, Y. H., LE, J., HERSEY, A., LUSCOMBE, C. N., REYNOLDS, D. P., BECK, G., SHERBORNE, B. & COOPER, I. 2002. On the mechanism of human intestinal absorption. *European Journal of Medicinal Chemistry,* 37**,** 595-605.

AGATONOVIC-KUSTRIN, S., BERESFORD, R. & YUSOF, A. P. M. 2001. Theoretically-derived molecular descriptors important in human intestinal absorption. *Journal of Pharmaceutical and Biomedical Analysis,* 25**,** 227-237.

ALI, J., CAMILLERI, P., BROWN, M. B., HUTT, A. J. & KIRTON, S. B. 2012. Revisiting the General Solubility Equation: In Silico Prediction of Aqueous Solubility Incorporating the Effect of Topographical Polar Surface Area. *Journal of Chemical Information and Modeling,* 52**,** 420-428.

AMIDON, G. L., LENNERNAS, H., SHAH, V. P. & CRISON, J. R. 1995. A theoretical basis for a biopharmaceutic drug classification - The correlation of in-vitro drug product dissolution and in-vivo bioavalability. *Pharmaceutical Research,* 12**,** 413-420.

ANDREWS, C. W., BENNETT, L. & YU, L. X. 2000. Predicting human oral bioavailability of a compound: Development of a novel quantitative structure-bioavailability relationship. *Pharmaceutical Research,* 17**,** 639-644.

APTULA, A. O., JELIAZKOVA, N. G., SCHULTZ, T. W. & CRONIN, M. T. D. 2005. The better predictive model: High q(2) for the training set or low root mean square error of prediction for the test set? *QSAR & Combinatorial Science,* 24**,** 385-396.

ARNOTT, J. A. & PLANEY, S. L. 2012. The influence of lipophilicity in drug discovery and design. *Expert Opinion on Drug Discovery,* 7**,** 863-75.

ARTURSSON, P. 1990. Epithelial transport of drugs in cell-culture.1. A model for studying the passive diffusion of drugs over intestinal absorptive (Caco-2) cells. *Journal of Pharmaceutical Sciences,* 79**,** 476-482.

ARTURSSON, P. & KARLSSON, J. 1991. Correlation between oral-drug absorption in humans and apparent drug perrmeability coefficients in human intestinal epithelial (Caco-2) cells. *Biochemical and Biophysical Research Communications,* 175**,** 880-885.

ASHFORD, M. 2007. Part 4: Biopharmaceutical principles of drug delivery. *In:* AULTON, M. E. (ed.) *Aulton's Pharmaceutics, The design and manufacture of medicines*. 3 ed. Philadelphia: Churchill Livingstone Elsevier.

AVDEEF, A. & TAM, K. Y. 2010. How Well Can the Caco-2/Madin-Darby Canine Kidney Models Predict Effective Human Jejunal Permeability? *Journal of Medicinal Chemistry,* 53**,** 3566-3584.

BAI, J. P. F., UTIS, A., CRIPPEN, G., HE, H. D., FISCHER, V., TULLMAN, R., YIN, H. Q., HSU, C. P., JIANG, L. & HWANG, K. K. 2004. Use of classification regression tree in predicting oral absorption in humans. *Journal of Chemical Information and Computer Sciences,* 44**,** 2061-2069.

BALABAN, A. T. 1982. Highly discriminating distance-based topological index. *Chemical Physics Letters,* 89**,** 399-404.

BALDI, P., BRUNAK, S., CHAUVIN, Y., ANDERSEN, C. A. F. & NIELSEN, H. 2000. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics,* 16**,** 412-424.

BALIMANE, P. V., CHONG, S. H. & MORRISON, R. A. 2000. Current methodologies used for evaluation of intestinal permeability and absorption. *Journal of Pharmacological and Toxicological Methods,* 44**,** 301-312.

BARANDELA, R., VALDOVINOS, R., SÁNCHEZ, J. S. & FERRI, F. 2004. The Imbalanced Training Sample Problem: Under or over Sampling? *In:* FRED, A.,

CAELLI, T., DUIN, R. W., CAMPILHO, A. & DE RIDDER, D. (eds.) *Structural, Syntactic, and Statistical Pattern Recognition*. Springer Berlin Heidelberg.

BERGSTROM, C. A. 2005. In silico predictions of drug solubility and permeability: two rate-limiting barriers to oral drug absorption. *Basic & Clinical Pharmacology & Toxicology,* 96**,** 156-61.

BERGSTROM, C. A., ANDERSSON, S. B., FAGERBERG, J. H., RAGNARSSON, G. & LINDAHL, A. 2014. Is the full potential of the biopharmaceutics classification system reached? *Eur J Pharm Sci,* 57**,** 224-31.

BERGSTROM, C. A. S., NORINDER, U., LUTHMAN, K. & ARTURSSON, P. 2002. Experimental and computational screening models for prediction of aqueous drug solubility. *Pharmaceutical Research,* 19**,** 182-188.

BERGSTROM, C. A. S., STRAFFORD, M., LAZOROVA, L., AVDEEF, A., LUTHMAN, K. & ARTURSSON, P. 2003. Absorption classification of oral drugs based on molecular surface properties. *Journal of Medicinal Chemistry,* 46**,** 558-570.

BERGSTROM, C. A. S., WASSVIK, C. M., NORINDER, U., LUTHMAN, K. & ARTURSSON, P. 2004. Global and Local Computational Models for Aqueous Solubility Prediction of Drug-Like Molecules. *Journal of Chemical Information and Computer Sciences,* 44**,** 1477-1488.

BLAGUS, R. & LUSA, L. 2010. Class prediction for high-dimensional class-imbalanced data. *Bmc Bioinformatics,* 11.

BNF 2012. *British National Formulary,* London, BMJ Group and Pharmaceutical Press.

BRANDSCH, M., KNUTTER, I. & BOSSE-DOENECKE, E. 2008. Pharmaceutical and pharmacological importance of peptide transporters. *Journal of Pharmacy and Pharmacology,* 60**,** 543-585.

BRAUN, A., HAMMERLE, S., SUDA, K., ROTHEN-RUTISHAUSER, B., GUNTHERT, M., KRAMER, S. D. & WUNDERLI-ALLENSPACH, H. 2000. Cell cultures as tools in biopharmacy. *European Journal of Pharmaceutical Sciences,* 11**,** S51-S60.

BREIMAN, L. 2001. Random forests. *Machine Learning,* 45**,** 5-32.

BREIMAN, L., FRIEDMAN, J., STONE, C. J. & OLSHEN, R. A. 1984. *Classification and Regression Trees,* Boca Raton, Chapman and Hall/CRC.

BRISKEANDERSON, M. J., FINLEY, J. W. & NEWMAN, S. M. 1997. The influence of culture time and passage number on the morphological and physiological development of Caco-2 cells. *Proceedings of the Society for Experimental Biology and Medicine,* 214**,** 248-257.

BROCCATELLI, F., CRUCIANI, G., BENET, L. Z. & OPREA, T. I. 2012. BDDCS Class Prediction for New Molecular Entities. *Molecular Pharmaceutics,* 9**,** 570-580.

BRODIE, B. B., SHORE, P. A. & HOGBEN, C. A. M. 1957. The Gastric Secretion Of Drugs: A pH Partition Hypothesis. *Journal of. Pharmacology and Experimental Therapeutics,* 119**,** 361-369.

BUCKLEY, S. T., FISCHER, S. M., FRICKER, G. & BRANDL, M. 2012. In vitro models to evaluate the permeability of poorly soluble drug entities: Challenges and perspectives. *European Journal of Pharmaceutical Sciences,* 45**,** 235-250.

BUNNAGE, M. E. 2011. Getting pharmaceutical R&D back on target. *Nature Chemical Biology,* 7**,** 335-339.

BURTON, P. S., GOODWIN, J. T., VIDMAR, T. J. & AMORE, B. M. 2002. Predicting drug absorption: How nature made it a difficult problem. *Journal of Pharmacology and Experimental Therapeutics,* 303**,** 889-895.

BUTLER, J. M. & DRESSMAN, J. B. 2010. The developability classification system: Application of biopharmaceutics concepts to formulation development. *Journal of Pharmaceutical Sciences,* 99**,** 4940-4954.

CALDWELL, G. W., RITCHIE, D. M., MASUCCI, J. A., HAGEMAN, W. & YAN, Z. 2001. The New Pre-Preclinical Paradigm: Compound Optimization in Early and Late Phase Drug Discovery. *Current Topics in Medicinal Chemistry,* 1**,** 353-366.

CALDWELL, G. W., YAN, Z. Y., TANG, W. M., DASGUPTA, M. & HASTING, B. 2009. ADME Optimization and Toxicity Assessment in Early- and Late-Phase Drug Discovery. *Current Topics in Medicinal Chemistry,* 9**,** 965-980.

CALDWELL, J., GARDNER, I. & SWALES, N. 1995. An Introduction To Drug Disposition - The Basic Principles Of Absorption, Distribution, Metabolism, And Excretion. *Toxicologic Pathology,* 23**,** 102-114.

CAO, D., WANG, J., ZHOU, R., LI, Y., YU, H. & HOU, T. 2012. ADMET evaluation in drug discovery. 11. PharmacoKinetics Knowledge Base (PKKB): a comprehensive database of pharmacokinetic and toxic properties for drugs. *Journal of Chemical Information and Modeling,* 52**,** 1132-7.

CARVALHO, A. P. L. F. & FREITAS, A. 2009. A Tutorial on Multi-label Classification Techniques. *In:* ABRAHAM, A., HASSANIEN, A.-E. & SNÁŠEL, V. (eds.) *Foundations of Computational Intelligence Volume 5*. p 177-195, Berlin, Springer.

CDER/FDA 2000. Waiver of In Vivo Bioavailability and Bioequivalence Studies for Immediate-Release Solid Oral Dosage Forms Based on a Biopharmaceutics Classification System, U.S. Department of Health and Human Services - Center for Drug Evaluation and Research. *Guidance for Industry*.

CERRI, R., DE CARVALHO, A. C. P. & FREITAS, A. A. 2011. Adapting non-hierarchical multilabel classification methods for hierarchical multilabel classification. *Intelligent Data Analysis,* 15**,** 861-887.

CHAN, L. M. S., LOWES, S. & HIRST, B. H. 2004. The ABCs of drug transport in intestine and liver: efflux proteins limiting drug absorption and bioavailability. *European Journal of Pharmaceutical Sciences,* 21**,** 25-51.

CHAN, O. H. & STEWART, B. H. 1996. Physicochemical and drug-delivery considerations for oral drug bioavailability. *Drug Discovery Today,* 1**,** 461-473.

CHAPELLE, O., SCHÖLKOPF, B. & ZIEN, A. 2006. *Semi-supervised learning*, MIT press Cambridge.

CHARMAN, W. N., PORTER, C. J. H., MITHANI, S. & DRESSMAN, J. B. 1997. Physicochemical and physiological mechanisms for the effects of food on drug absorption: The role of lipids and pH. *Journal of Pharmaceutical Sciences,* 86**,** 269-282.

CHEN, X. Q., CHO, S. J., LI, Y. & VENKATESH, S. 2002. Prediction of aqueous solubility of organic compounds using a quantitative structure-property relationship. *Journal of Pharmaceutical Sciences,* 91**,** 1838-1852.

CHENG, A. L. & MERZ, K. M. 2003. Prediction of aqueous solubility of a diverse set of compounds using quantitative structure-property relationships. *Journal of Medicinal Chemistry,* 46**,** 3572-3580.

CHERKASOV, A., MURATOV, E. N., FOURCHES, D., VARNEK, A., BASKIN, II, CRONIN, M., DEARDEN, J., GRAMATICA, P., MARTIN, Y. C., TODESCHINI, R., CONSONNI, V., KUZ'MIN, V. E., CRAMER, R., BENIGNI, R., YANG, C., RATHMAN, J., TERFLOTH, L., GASTEIGER, J., RICHARD, A. & TROPSHA, A. 2014. QSAR modeling: where have you been? Where are you going to? *Journal of Medicinal Chemistry,* 57**,** 4977-5010.

CHIOU, W. L. 2001. The rate and extent of oral bioavailability versus the rate and extent of oral absorption: Clarification and recommendation of terminology. *Journal of Pharmacokinetics and Pharmacodynamics,* 28**,** 3-6.

CHU, K. A. & YALKOWSKY, S. H. 2009. An interesting relationship between drug absorption and melting point. *International Journal of Pharmaceutics,* 373**,** 24-40.

CHUASUWAN, B., BINJESOH, V., POLLI, J., ZHANG, H., AMIDON, G., JUNGINGER, H., MIDHA, K., SHAH, V., STAVCHANSKY, S. & DRESSMAN, J. 2009. Biowaiver monographs for immediate release solid oral dosage forms: Diclofenac sodium and diclofenac potassium. *Journal of Pharmaceutical Sciences,* 98**,** 1206-1219.

CLARE, A. & KING, R. 2001. Knowledge Discovery in Multi-label Phenotype Data. *In:* RAEDT, L. & SIEBES, A. (eds.) *Principles of Data Mining and Knowledge Discovery*. Springer Berlin Heidelberg.

CLARK, D. E. 1999. Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 1. Prediction of intestinal absorption. *Journal of Pharmaceutical Sciences,* 88**,** 807-814.

COHEN, J. 1968. Weighed kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin,* 70.

COMER, J. E. A. 2003. High-throughput Measurement of log D and pka. *In:* WATERBEEMD, H. V. D., LENNERNÄS, H., ARTURSSON, P., MANNHOLD, R., KUBINYI, H. & FOLKERS, G. (eds.) *Drug Bioavailability: Estimation of Solubility, Permeability, Absorption and Bioavailability (Methods and Principles in Medicinal Chemistry)*. 1 ed. Weinheim: Wiley-VCH.

COOPER, J. A., SARACCI, R. & COLE, P. 1979. DESCRIBING THE VALIDITY OF CARCINOGEN SCREENING-TESTS. *British Journal of Cancer,* 39**,** 87-89.

CORREA, E. S., FREITAS, A. A. & JOHNSON, C. G. 2008. Particle swarm for attribute selection in Bayesian classification: an application to protein function prediction. *Journal of Artificial Evolution and Applications,* 2008**,** 1-12.

COSTA, E., LORENA, A., CARVALHO, A. P. L. F., FREITAS, A. & HOLDEN, N. 2007. Comparing Several Approaches for Hierarchical Classification of Proteins with Decision Trees. *In:* SAGOT, M.-F. & WALTER, M. T. (eds.) *Advances in Bioinformatics and Computational Biology*. Springer Berlin Heidelberg.

CRISON, J. R., TIMMINS, P., KEUNG, A., UPRETI, V. V., BOULTON, D. W. & SCHEER, B. J. 2012. Biowaiver approach for biopharmaceutics classification system class 3 compound metformin hydrochloride using in silico modeling. *Journal of Pharmaceutical Sciences,* 101**,** 1773-82.

CRUCIANI, G., PASTOR, M. & GUBA, W. 2000. VolSurf: a new tool for the pharmacokinetic optimization of lead compounds. *European Journal of Pharmaceutical Sciences,* 11**,** S29-S39.

CUMMINGS, D., J. 2006. Pharmaceutical Drug Discovery: Designing the Blockbuster Drug. *In:* DEAN, A. & LEWIS, S. (eds.) *Screening Methods for Experimentation in Industry, Drug Discovery, and Genetics*. 1 ed. New York: Springer.

CUSTODIO, J. M., WU, C.-Y. & BENET, L. Z. 2008. Predicting drug disposition, absorption/elimination/transporter interplay and the role of food on drug absorption. *Advanced Drug Delivery Reviews,* 60**,** 717-733.

CZODROWSKI, P. 2013. hERG Me Out. *Journal of Chemical Information and Modeling,* 53**,** 2240-2251.

DAHAN, A., MILLER, J. & AMIDON, G. 2009. Prediction of Solubility and Permeability Class Membership: Provisional BCS Classification of the World's Top Oral Drugs. *The AAPS Journal,* 11**,** 740-746.

DARWICH, A. S., NEUHOFF, S., JAMEI, M. & ROSTAMI-HODJEGAN, A. 2010. Interplay of Metabolism and Transport in Determining Oral Drug Absorption and Gut Wall Metabolism: A Simulation Assessment Using the "Advanced Dissolution, Absorption, Metabolism (ADAM)" Model. *Current Drug Metabolism,* 11**,** 716-729.

DAVIES, N. M., TENG, X. W. & SKJODT, N. M. 2003. Pharmacokinetics of rofecoxib. *Clinical Pharmacokinetics,* 42**,** 545-556.

DAVIS, A. M. & BRUNEA, P. 2003. In Silico Prediction of Solubility. *In:* WATERBEEMD, H. V. D., LENNERNÄS, H., ARTURSSON, P., MANNHOLD, R., KUBINYI, H. & FOLKERS, G. (eds.) *Drug Bioavailability: Estimation of Solubility, Permeability, Absorption and Bioavailability (Methods and Principles in Medicinal Chemistry)*. 1 ed. Weinheim: Wiley-VCH.

DAVIS, S. S., HARDY, J. G. & FARA, J. W. 1986. Transit Of Pharmaceutical Dosage Forms Through The Small-Intestine. *Gut,* 27**,** 886-892.

DECONINCK, E., ATES, H., CALLEBAUT, N., VAN GYSEGHEM, E. & VANDER HEYDEN, Y. 2007. Evaluation of chromatographic descriptors for the prediction of gastro-intestinal absorption of drugs. *Journal of Chromatography A,* 1138**,** 190-202.

DECONINCK, E., HANCOCK, T., COOMANS, D., MASSART, D. L. & VANDER HEYDEN, Y. 2005. Classification of drugs in absorption classes using the classification and regression trees (CART) methodology. *Journal of Pharmaceutical and Biomedical Analysis,* 39**,** 91-103.

DELANEY, J. S. 2004. ESOL: Estimating Aqueous Solubility Directly from Molecular Structure. *Journal of Chemical Information and Computer Sciences,* 44**,** 1000-1005.

DI, L., ARTURSSON, P., AVDEEF, A., ECKER, G. F., FALLER, B., FISCHER, H., HOUSTON, J. B., KANSY, M., KERNS, E. H., KRÄMER, S. D., LENNERNÄS, H. & SUGANO, K. 2012. Evidence-based approach to assess passive diffusion and carrier-mediated drug transport. *Drug Discovery Today,* 17**,** 905-912.

DI, L. & KERNS, E. H. 2003. Profiling drug-like properties in discovery research. *Current Opinion in Chemical Biology,* 7**,** 402-408.

DI, L., WHITNEY-PICKETT, C., UMLAND, J. P., ZHANG, H., ZHANG, X., GEBHARD, D. F., LAI, Y. R., FEDERICO, J. J., DAVIDSON, R. E., SMITH, R., REYNER, E. L., LEE, C., FENG, B., ROTTER, C., VARMA, M. V., KEMPSHALL, S., FENNER, K., EL-KATTAN, A. F., LISTON, T. E. & TROUTMAN, M. D. 2011. Development of a New Permeability Assay Using Low-Efflux MDCKII Cells. *Journal of Pharmaceutical Sciences,* 100**,** 4974-4985.

DIETTERICH, T. 2000. Ensemble Methods in Machine Learning. *Multiple Classifier Systems*. Springer Berlin Heidelberg.

DIMASI, J. A., HANSEN, R. W. & GRABOWSKI, H. G. 2003. The price of innovation: new estimates of drug development costs. *Journal of Health Economics,* 22**,** 151-185.

DOBSON, P. D. & KELL, D. B. 2008. Opinion - Carrier-mediated cellular uptake of pharmaceutical drugs: an exception or the rule? *Nature Reviews Drug Discovery,* 7**,** 205-220.

DOWTY, M. E., MESSING, D. M., LAI, L. & KIRKOVSKY, L. 2011. ADME. *In:* TSAIOUN, K. & KATES, S. A. (eds.) *ADMET for Medicinal Chemists A Practical Guide*. New Jersey: Wiley.

DRESSMAN, J. B. 1986. Comparison Of Canine And Human Gastrointestinal Physiology. *Pharmaceutical Research,* 3**,** 123-131.

DRESSMAN, J. B., AMIDON, G. L. & FLEISHER, D. 1985. Absorption potential: Estimating the fraction absorbed for orally administered compounds. *Journal of Pharmaceutical Sciences,* 74**,** 588-589.

DUCHOWICZ, P. R., TALEVI, A., BRUNO-BLANCH, L. E. & CASTRO, E. A. 2008. New QSPR study for the prediction of aqueous solubility of drug-like compounds. *Bioorganic & Medicinal Chemistry,* 16**,** 7944-7955.

DUDEK, A. Z., ARODZ, T. & GALVEZ, J. 2006. Computational methods in developing quantitative structure-activity relationships (QSAR): A review. *Combinatorial Chemistry & High Throughput Screening,* 9**,** 213-228.

EAKINS, G. L., ALFORD, J. S., TIEGS, B. J., BREYFOGLE, B. E. & STEARMAN, C. J. 2011. Tuning HOMO–LUMO levels: trends leading to the design of 9-fluorenone scaffolds with predictable electronic and optoelectronic properties. *Journal of Physical Organic Chemistry,* 24**,** 1119-1128.

EGAN, W. J. & LAURI, G. 2002. Prediction of intestinal permeability. *Advanced Drug Delivery Reviews,* 54**,** 273-289.

EGAN, W. J., MERZ, K. M. & BALDWIN, J. J. 2000. Prediction of drug absorption using multivariate statistics. *Journal of Medicinal Chemistry,* 43**,** 3867-3877.

EITRICH, T., KLESS, A., DRUSKA, C., MEYER, W. & GROTENDORST, J. 2006. Classification of Highly Unbalanced CYP450 Data of Drugs Using Cost Sensitive

Machine Learning Techniques. *Journal of Chemical Information and Modeling,* 47**,** 92-103.
EKINS, S., WALLER, C. L., SWAAN, P. W., CRUCIANI, G., WRIGHTON, S. A. & WIKEL, J. H. 2000. Progress in predicting human ADME parameters in silico. *Journal of Pharmacological and Toxicological Methods,* 44**,** 251-272.
EKLUND, M., NORINDER, U., BOYER, S. & CARLSSON, L. 2014. Choosing Feature Selection and Learning Algorithms in QSAR. *Journal of Chemical Information and Modeling,* 54**,** 837-843.
EL-HENAWY, A. A., KHOWDIARY, M. M., BADAWI, A. B. & SOLIMAN, H. M. 2013. In Vivo Anti-Leukemia, Quantum Chemical Calculations and ADMET Investigations of Some Quaternary and Isothiouronium Surfactants. *Pharmaceuticals,* 6**,** 634-49.
EMA 2010. Guideline on the Investigation of Bioequivalence, Committee for Medicinal Products for Human Use (CHMP), http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2010/01/WC500070039.pdf [Accessed 11 September 2014]. *Committee for Medicinal Products for Human Use (CHMP),*.
ENGLUND, G., RORSMAN, F., RONNBLOM, A., KARLBOM, U., LAZOROVA, L., GRASJO, J., KINDMARK, A. & ARTURSSON, P. 2006. Regional levels of drug transporters along the human intestinal tract: Co-expression of ABC and SLC transporters and comparison with Caco-2 cells. *European Journal of Pharmaceutical Sciences,* 29**,** 269-277.
ERTL, P., ROHDE, B. & SELZER, P. 2000. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *Journal of Medicinal Chemistry,* 43**,** 3714-3717.
ESTUDANTE, M., MORAIS, J. G., SOVERAL, G. & BENET, L. Z. 2013. Intestinal drug transporters: an overview. *Advanced Drug Delivery Reviews,* 65**,** 1340-56.
FAGERHOLM, U. 2007. Prediction of human pharmacokinetics - gut-wall metabolism. *Journal of Pharmacy and Pharmacology,* 59**,** 1335-1343.
FISHER, M. B., PAINE, M. F., STRELEVITZ, T. J. & WRIGHTON, S. A. 2001. The role of hepatic and extrahepatic UDP-glucuronosyltransferases in human drug metabolism. *Drug Metabolism Reviews,* 33**,** 273-297.
FLEISHER, D., LI, C., ZHOU, Y., PAO, L. H. & KARIM, A. 1999. Drug, meal and formulation interactions influencing drug absorption after oral administration - Clinical implications. *Clinical Pharmacokinetics,* 36**,** 233-254.
FOGH, J. & TREMPE, G. 1975. Human Tumor Cells In Vitro *In:* FOGH, J. (ed.) *Human Tumor Cells In Vitro*. 1 ed. New York: Plenum Press.
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 Modeling,* 50**,** 1189-1204.
FREIRE, A. C., BASIT, A. W., CHOUDHARY, R., PIONG, C. W. & MERCHANT, H. A. 2011. Does sex matter? The influence of gender on gastrointestinal physiology and drug delivery. *International Journal of Pharmaceutics,* 415**,** 15-28.
GAO, Z. 2012. Development of a Continuous Dissolution/Absorption System—a Technical Note. *AAPS PharmSciTech,* 13**,** 1287-1292.
GASTEIGER, J. & MARSILI, M. 1980. Iterative partial equalization of orbital electronegativity - a rapid access to atomic charges. *Tetrahedron,* 36**,** 3219-3228.
GAZPIO, C., SÁNCHEZ, M., GARCÍA-ZUBIRI, I. X., VÉLAZ, I., MARTÍNEZ-OHÁRRIZ, C., MARTÍN, C. & ZORNOZA, A. 2005. HPLC and solubility study of the interaction between pindolol and cyclodextrins. *Journal of Pharmaceutical and Biomedical Analysis,* 37**,** 487-492.
GEERTS, T. & HEYDEN, Y. V. 2011. In Silico Predictions of ADME-Tox Properties: Drug Absorption. *Combinatorial Chemistry & High Throughput Screening,* 14**,** 339-361.

GERTZ, M., HARRISON, A., HOUSTON, J. B. & GALETIN, A. 2010. Prediction of Human Intestinal First-Pass Metabolism of 25 CYP3A Substrates from In Vitro Clearance and Permeability Data. *Drug Metabolism and Disposition,* 38**,** 1147-1158.

GHAFOURIAN, T. & BOZORGI, A. H. A. 2010. Estimation of drug solubility in water, PEG 400 and their binary mixtures using the molecular structures of solutes. *European Journal of Pharmaceutical Sciences,* 40**,** 430-440.

GHAFOURIAN, T. & CRONIN, M. T. D. 2005. The impact of variable selection on the modelling of oestrogenicity. *SAR and QSAR in Environmental Research,* 16**,** 171-190.

GHAFOURIAN, T., NEWBY, D. & FREITAS, A. A. 2012. The impact of training set data distributions for modelling of passive intestinal absorption. *International Journal of Pharmaceutics,* 436**,** 711-720.

GHIBELLINI, G., LESLIE, E. M. & BROUWER, K. L. R. 2006. Methods to evaluate biliary excretion of drugs in humans: An updated review. *Molecular Pharmaceutics,* 3**,** 198-211.

GIACOMINI, K. M., HUANG, S. M., TWEEDIE, D. J., BENET, L. Z., BROUWER, K. L. R., CHU, X. Y., 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. & INTERNATIONAL, T. 2010. Membrane transporters in drug development. *Nature Reviews Drug Discovery,* 9**,** 215-236.

GIDAL, B. E. 2006. Drug absorption in the elderly: Biopharmaceutical considerations for the antiepileptic drugs. *Epilepsy Research,* 68**,** S65-S69.

GINI, G., CRACIUN, M. V., KONIG, C. & BENFENATI, E. 2004. Combining unsupervised and supervised artificial neural networks to predict aquatic toxicity. *Journal of Chemical Information and Computer Sciences,* 44**,** 1897-1902.

GLATT, H., BOEING, H., ENGELKE, C. E. H., KUHLOW, L. M. A., PABEL, U., POMPLUN, D., TEUBNER, W. & MEINL, W. 2001. Human cytosolic sulphotransferases: genetics, characteristics, toxicological aspects. *Mutation Research-Fundamental and Molecular Mechanisms of Mutagenesis,* 482**,** 27-40.

GLEESON, M. P., 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.

GLEESON, M. P., MODI, S., BENDER, A., ROBINSON, R. L., KIRCHMAIR, J., PROMKATKAEW, M., HANNONGBUA, S. & GLEN, R. C. 2012. The challenges involved in modeling toxicity data in silico: a review. *Current Pharmaceutical Design,* 18**,** 1266-91.

GOLBRAIKH, A., MURATOV, E., FOURCHES, D. & TROPSHA, A. 2014. Data Set Modelability by QSAR. *Journal of Chemical Information and Modeling,* 54**,** 1-4.

GOLBRAIKH, A. & TROPSHA, A. 2002a. Beware of q(2)! *Journal of Molecular Graphics & Modelling,* 20**,** 269-276.

GOLBRAIKH, A. & TROPSHA, A. 2002b. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *Journal of Computer-Aided Molecular Design,* 16**,** 357-369.

GOLDBERG, D. E. 1989. *Genetic algorithms in search, optimization, and machine learning*, Boston, Addison-Wesley Longman Publishing.

GONÇALVES, E. C., PLASTINO, A. & FREITAS, A. A. 2013. A Genetic Algorithm for Optimizing the Label Ordering in Multi-Label Classifier Chains. *In:* Proceedings of the 2013 25th IEEE International Conference on Tools with Artificial Intelligence (ICTAI), 2013. IEEE Computer Society Conference Publishing Services (CPS), 469-476.

GOODARZI, M., DEJAEGHER, B. & VANDER HEYDEN, Y. 2012. Feature Selection Methods in QSAR Studies. *Journal of AOAC International,* 95**,** 636-651.

GOZALBES, R., JACEWICZ, M., ANNAND, R., TSAIOUN, K. & PINEDA-LUCENA, A. 2011. QSAR-based permeability model for drug-like compounds. *Bioorganic & Medicinal Chemistry,* 19**,** 2615-2624.

GOZALBES, R. & PINEDA-LUCENA, A. 2010. QSAR-based solubility model for drug-like compounds. *Bioorganic & Medicinal Chemistry,* 18**,** 7078-7084.

GRAMATICA, P. 2007. Principles of QSAR models validation: internal and external. *QSAR & Combinatorial Science,* 26**,** 694-701.

GRES, M. C., JULIAN, B., BOURRIE, M., MEUNIER, V., ROQUES, C., BERGER, M., BOULENC, X., BERGER, Y. & FABRE, G. 1998. Correlation between oral drug absorption in humans, and apparent drug permeability in TC-7 cells, a human epithelial intestinal cell line: Comparison with the parental Caco-2 cell line. *Pharmaceutical Research,* 15**,** 726-733.

GRISONI, F., CASSOTTI, M. & TODESCHINI, R. 2014. Reshaped Sequential Replacement for variable selection in QSPR: comparison with other reference methods. *Journal of Chemometrics,* 28**,** 249-259.

GUBBINS, P. & BERTCH, K. 1991. Drug Absorption in Gastrointestinal Disease and Surgery. *Clinical Pharmacokinetics,* 21**,** 431-447.

GUYON, I. & ELISSEEFF, A. 2003. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research,* 3**,** 1157-1182.

HAHN, G. J. 1977. Fitting Regression Models with No Intercept Term. *Journal of Quality Technology,* 9**,** 56-61.

HALL, L. H. & KIER, L. B. 1991. The Molecular Connectivity Chi Indices and Kappa Shape Indices in Structure-Property Modeling. *In:* BOYD, D. & LIPKOWITZ, K. (eds.) *Reviews in Computational Chemistry*. New York: VCH.

HALL, M., FRANK, E., HOLMES, G., PFAHRINGER, B., REUTEMANN, P. & WITTEN, I. H. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations,* 11**,** 10-18.

HAMMETT, L., P. 1935. Some Relations between Reaction Rates and Equilibrium Constants. *Chemical Reviews,* 17**,** 125-136.

HAMMETT, L., P. 1970. *Physical Organic Chemistry* New York, McGraw-Hill.

HANSCH, C., MALONEY, P., P., FUJITA, T. & MUIR, M., R. 1962. Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature,* 194**,** 178-180.

HAWKINS, D. M. 2004. The problem of overfitting. *Journal of Chemical Information and Computer Sciences,* 44**,** 1-12.

HAWKINS, D. M., BASAK, S. C. & MILLS, D. 2003. Assessing model fit by cross-validation. *Journal of Chemical Information and Computer Sciences,* 43**,** 579-586.

HAYESHI, R., HILGENDORF, C., ARTURSSON, P., AUGUSTIJNS, P., BRODIN, B., DEHERTOGH, P., FISHER, K., FOSSATI, L., HOVENKAMP, E., KORJAMO, T., MASUNGI, C., MAUBON, N., MOLS, R., MULLERTZ, A., MONKKONEN, J., O'DRISCOLL, C., OPPERS-TIEMISSEN, H. M., RAGNARSSON, E. G. E., ROOSEBOOM, M. & UNGELL, A. L. 2008. Comparison of drug transporter gene expression and functionality in Caco-2 cells from 10 different laboratories. *European Journal of Pharmaceutical Sciences,* 35**,** 383-396.

HEADING, R., NIMMO, J., PRESCOTT, L. & TOTHILL, P. 1973. The dependence of paracetamol absorption on the rate of gastric emptying. *British Journal of Pharmacology,* 47**,** 415-421.

HEBERT, M. F. 1997. Contributions of hepatic and intestinal metabolism and P-glycoprotein to cyclosporine and tacrolimus oral drug delivery. *Advanced Drug Delivery Reviews,* 27**,** 201-214.

HEWITT, M., CRONIN, M. T. D., ENOCH, S. J., MADDEN, J. C., ROBERTS, D. W. & DEARDEN, J. C. 2009. In Silico Prediction of Aqueous Solubility: The Solubility Challenge. *Journal of Chemical Information and Modeling,* 49**,** 2572-2587.

HIDALGO, I. J., RAUB, T. J. & BORCHARDT, R. T. 1989. Characterization of the human colon carcinoma cell line (Caco-2) as a model system for intestinal epithelial permeability. *Gastroenterology,* 96**,** 736-739.

HILGENDORF, C., AHLIN, G., SEITHEL, A., ARTURSSON, P., UNGELL, A. L. & KARLSSON, J. 2007. Expression of thirty-six drug transporter genes in human intestine, liver, kidney, and organotypic cell lines. *Drug Metabolism and Disposition,* 35**,** 1333-1340.

HILGENDORF, C., SPAHN-LANGGUTH, H., REGARDH, C. G., LIPKA, E., AMIDON, G. L. & LANGGUTH, P. 2000. Caco-2 versus Caco-2/HT29-MTX co-cultured cell lines: Permeabilities via diffusion, inside- and outside-directed carrier-mediated transport. *Journal of Pharmaceutical Sciences,* 89**,** 63-75.

HOLLAND, J. H. 1975. *Adaptation in Natural and Artificial Systems,* Michigan, University of Michigan Press (re-issued by MIT Press 1992).

HOU, T. J., LI, Y. Y., ZHANG, W. & WANG, J. M. 2009. Recent Developments of In Silico Predictions of Intestinal Absorption and Oral Bioavailability. *Combinatorial Chemistry & High Throughput Screening,* 12**,** 497-506.

HOU, T. J., WANG, J. M. & LI, Y. Y. 2007a. ADME evaluation in drug discovery. 8. The prediction of human intestinal absorption by a support vector machine. *Journal of Chemical Information and Modeling,* 47**,** 2408-2415.

HOU, T. J., WANG, J. M., ZHANG, W., WANG, W. & XU, X. 2006. Recent advances in computational prediction of drug absorption and permeability in drug discovery. *Current Medicinal Chemistry,* 13**,** 2653-2667.

HOU, T. J., WANG, J. M., ZHANG, W. & XU, X. J. 2007b. ADME evaluation in drug discovery. 6. Can oral bioavailability in humans be effectively predicted by simple molecular property-based rules? *Journal of Chemical Information and Modeling,* 47**,** 460-463.

HOU, T. J., WANG, J. M., ZHANG, W. & XU, X. J. 2007c. ADME evaluation in drug discovery. 7. Prediction of oral absorption by correlation and classification. *Journal of Chemical Information and Modeling,* 47**,** 208-218.

IRVINE, J. D., TAKAHASHI, L., LOCKHART, K., CHEONG, J., TOLAN, J. W., SELICK, H. E. & GROVE, J. R. 1999. MDCK (Madin-Darby canine kidney) cells: A tool for membrane permeability screening. *Journal of Pharmaceutical Sciences,* 88**,** 28-33.

JACOBSEN, W., KIRCHNER, G., HALLENSLEBEN, K., MANCINELLI, L., DETERS, M., HACKBARTH, I., BANER, K., BENET, L. Z., SEWING, K. F. & CHRISTIANS, U. 1999. Small intestinal metabolism of the 3-hydroxy-3-methylglutary-coenzyme A reductase inhibitor lovastatin and comparison with pravastatin. *Journal of Pharmacology and Experimental Therapeutics,* 291**,** 131-139.

JAIN, N. & YALKOWSKY, S. H. 2001. Estimation of the aqueous solubility I: Application to organic nonelectrolytes. *Journal of Pharmaceutical Sciences,* 90**,** 234-252.

JAIN, P. & YALKOWSKY, S. H. 2010. Prediction of aqueous solubility from SCRATCH. *International Journal of Pharmaceutics,* 385**,** 1-5.

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,* 33**,** 445-59.

JENSEN, B. F., REFSGAARD, H. H. F., BRO, R. & BROCKHOFF, P. B. 2005. Classification of membrane permeability of drug candidates: A methodological investigation. *QSAR & Combinatorial Science,* 24**,** 449-457.

JONKMAN, J. H. G. & HUNT, C. A. 1983. ION-PAIR ABSORPTION OF IONIZED DRUGS - FACT OR FICTION. *Pharmaceutisch Weekblad-Scientific Edition,* 5**,** 41-48.

JUNGJIT, S., FREITAS, A. A., MICHAELIS, M. & CINATL, J. 2013. Two Extensions to Multi-label Correlation-Based Feature Selection: A Case Study in Bioinformatics. *In:* Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on, 13-16 Oct 2013. 1519-1524.

KASIM, N. A., WHITEHOUSE, M., RAMACHANDRAN, C., BERMEJO, M., LENNERNÄS, H., HUSSAIN, A. S., JUNGINGER, H. E., STAVCHANSKY, S. A., MIDHA, K. K. & SHAH, V. P. 2004. Molecular properties of WHO essential drugs and provisional biopharmaceutical classification. *Molecular Pharmaceutics,* 1**,** 85-96.

KAY, K. 2011. Introduction. *In:* TSAIOUN, K. & KATES, S. A. (eds.) *ADMET for Medicinal Chemists A Practical Guide*. New Jersey: Wiley.

KELL, D. B., DOBSON, P. D., BILSLAND, E. & OLIVER, S. G. 2013. The promiscuous binding of pharmaceutical drugs and their transporter-mediated uptake into cells: what we (need to) know and how we can do so. *Drug Discovery Today,* 18**,** 218-239.

KENNARD, R. W. & STONE, L. A. 1969. Computer Aided Design of Experiments. *Technometrics,* 11**,** 137-148.

KERNS, E. H. & DI, L. 2008. *Drug like properties: Concepts, Structure Design and Methods from ADME to Toxicity Optimisation,* Burlington, Academic Press Elsevier.

KHANDELWAL, A., BAHADDURI, P. M., CHANG, C., POLLI, J. E., SWAAN, P. W. & EKINS, S. 2007. Computational models to assign biopharmaceutics drug disposition classification from molecular structure. *Pharmaceutical Research,* 24**,** 2249-2262.

KHAZAEINIA, T., RAMSEY, A. A. & TAM, Y. K. 2000. The effects of exercise on the pharmacokinetics of drugs. *Journal of Pharmacy and Pharmaceutical Sciences,* 3**,** 292-302.

KIKUCHI, A., TOMOYASU, T., TANAKA, M., KANAMITSU, K., SASABE, H., MAEDA, T., ODOMI, M. & TAMAI, I. 2009. Peptide derivation of poorly absorbable drug allows intestinal absorption via peptide transporter. *Journal of Pharmaceutical Sciences,* 98**,** 1775-87.

KING, R. S. 2009. Biotransformations in Drug Metabolism. *In:* NASSAR, A. F., HOLLENBERG, P. F. & SCATINA, J. (eds.) *Drug Metabolism Handbook Concepts and Applications*. 1 ed. New Jersey: Wiley.

KIRKOVSKY, L. & ZUTSHI, A. 2011. Pharmacokinetics for Medicinal Chemists. *In:* TSAIOUN, K. & KATES, A. S. (eds.) *ADMET for Medicinal Chemists A Practical Guide*. 1 ed. New Jersey: Wiley.

KITTLER, J. 1978. Feature set search algorithms. *In:* CHEN, C. H. (ed.) *Pattern Recognition and Signal Processing*. The Netherlands: Sijthoff and Noordhoff,.

KLOPMAN, G., STEFAN, L. R. & SAIAKHOV, R. D. 2002. ADME evaluation 2. A computer model for the prediction of intestinal absorption in humans. *European Journal of Pharmaceutical Sciences,* 17**,** 253-263.

KOHAVI, R. & JOHN, G. H. 1997. Wrappers for feature subset selection. *Artificial Intelligence,* 97**,** 273-324.

KOLA, I. & LANDIS, J. 2004. Can the pharmaceutical industry reduce attrition rates? *Nature Reviews Drug Discovery,* 3**,** 711-715.

KOREN, G. 1997. Therapeutic drug monitoring principles in the neonate. *Clinical Chemistry,* 43**,** 222-227.

KORTAGERE, S. & EKINS, S. 2010. Troubleshooting computational methods in drug discovery. *Journal of Pharmacological and Toxicological Methods,* 61**,** 67-75.

KRISHNASWAMY, S., DUAN, S. X., VON MOLTKE, L. L., GREENBLATT, D. J. & COURT, M. H. 2003. Validation of serotonin (5-hydroxtryptamine) as an in vitro substrate probe for human UDP-glucuronosyltransferase (UGT) 1A6. *Drug Metabolism and Disposition,* 31**,** 133-139.

KU, M. S. 2008. Use of the biopharmaceutical classification system in early drug development. *AAPS Journal,* 10**,** 208-212.

KWON, Y. 2002. *Handbook of Essential Pharmacokinetics, Pharmacodynamics, and Drug Metabolism for Industrial Scientists,* New York, Springer.

LAURIA, A., IPPOLITO, M. & ALMERICO, A. M. 2009. Combined Use of PCA and QSAR/QSPR to Predict the Drugs Mechanism of Action. An Application to the NCI ACAM Database. *QSAR & Combinatorial Science,* 28**,** 387-395.

LE FERREC, E., CHESNE, C., ARTUSSON, P., BRAYDEN, D., FABRE, G., GIRES, P., GUILLOU, F., ROUSSET, M., RUBAS, W. & SCARINO, M. L. 2001. In vitro models of the intestinal barrier - The report and recommendations of ECVAM Workshop 46. *Atla-Alternatives to Laboratory Animals,* 29**,** 649-668.

LEESON, P. D. & SPRINGTHORPE, B. 2007. The influence of drug-like concepts on decision-making in medicinal chemistry. *Nature Reviews Drug Discovery,* 6**,** 881-890.

LENNERNAS, H. 1997. Human jejunal effective permeability and its correlation with preclinical drug absorption models. *Journal of Pharmacy and Pharmacology,* 49**,** 627-638.

LENNERNAS, H. 2007. Animal data: The contributions of the Ussing Chamber and perfusion systems to predicting human oral drug delivery in vivo. *Advanced Drug Delivery Reviews,* 59**,** 1103-1120.

LENNERNAS, H. & ABRAHAMSSON, B. 2005. The use of biopharmaceutic classification of drugs in drug discovery and development: current status and future extension. *Journal of Pharmacy and Pharmacology,* 57**,** 273-285.

LI, G.-Z., YANG, J. Y., LU, W.-C. & LI, D. 2008. Improving prediction accuracy of drug activities by utilising unlabelled instances with feature selection. *International Journal of Computational Biology and Drug Design,* 1**,** 1-13.

LIN, J. H., CHIBA, M. & BAILLIE, T. A. 1999. Is the role of the small intestine in first-pass metabolism overemphasized? *Pharmacological Reviews,* 51**,** 135-157.

LINDENBERG, M., KOPP, S. & DRESSMAN, J. B. 2004. Classification of orally administered drugs on the World Health Organization Model list of Essential Medicines according to the biopharmaceutics classification system. *European Journal of Pharmaceutics and Biopharmaceutics,* 58**,** 265-278.

LIPINSKI, C. A. 2000. Drug-like properties and the causes of poor solubility and poor permeability. *Journal of Pharmacological and Toxicological Methods,* 44**,** 235-249.

LIPINSKI, C. A., LOMBARDO, F., DOMINY, B. W. & FEENEY, P. J. 1997. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews,* 23**,** 3-25.

LIU, H. & SETIONO, R. 1995. Chi2: Feature selection and discretization of numeric attributes. *In:* VASSILOPOULOS, J. F. (ed.) *Seventh International Conference on Tools with Artificial Intelligence, Proceedings*.

LIU, Y. 2004. A comparative study on feature selection methods for drug discovery. *Journal of Chemical Information and Computer Sciences,* 44**,** 1823-1828.

LOBELL, M., HENDRIX, M., HINZEN, B., KELDENICH, J., MEIER, H., SCHMECK, C., SCHOHE-LOOP, R., WUNBERG, T. & HILLISCH, A. 2006. In silico ADMET traffic lights as a tool for the prioritization of HTS hits. *Chemmedchem,* 1**,** 1229-1236.

LOEBSTEIN, R., LALKIN, A. & KOREN, G. 1997. Pharmacokinetic changes during pregnancy and their clinical relevance. *Clinical Pharmacokinetics,* 33**,** 328-343.

MACHERAS, P. & KARALIS, V. 2014. A non-binary biopharmaceutical classification of drugs: the ABGamma system. *Int J Pharm,* 464**,** 85-90.

MAHALANOBIS, P. C. 1936. On the generalised distance in statistics. *In:* Proceedings of the National Institute of Sciences of India, 1936. 49-55.

MALO, N., HANLEY, J. A., CERQUOZZI, S., PELLETIER, J. & NADON, R. 2006. Statistical practice in high-throughput screening data analysis. *Nature Biotechnology,* 24**,** 167-175.

MARTIN, T. M., HARTEN, P., YOUNG, D. M., MURATOV, E. N., GOLBRAIKH, A., ZHU, H. & TROPSHA, A. 2012. Does Rational Selection of Training and Test Sets Improve the Outcome of QSAR Modeling? *Journal of Chemical Information and Modeling,* 52**,** 2570-2578.

MARTINDALE 2009. *Martindale The Complete Drug Reference,* London, Pharmaceutical Press.

MARTINEZ, M. N. & AMIDON, G. L. 2002. A mechanistic approach to understanding the factors affecting drug absorption: A review of fundamentals. *Journal of Clinical Pharmacology,* 42**,** 620-643.

MATSSON, P., BERGSTROM, C. A. S., NAGAHARA, N., TAVELIN, S., NORINDER, U. & ARTURSSON, P. 2005. Exploring the role of different drug transport routes in permeability screening. *Journal of Medicinal Chemistry,* 48**,** 604-613.

MATTHEWS, B. W. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta,* 405**,** 442-451.

MCCALLUM, A. 1999. Multi-label text classification with a mixture model trained by EM. *In:* AAAI'99 Workshop on Text Learning, 1999. 1-7.

MEHRAMIZI, A., ALIJANI, B., POURFARZIB, M., DORKOOSH, F. A. & RAFIEE – TEHRANI, M. 2007. Solid Carriers for Improved Solubility of Glipizide in Osmotically Controlled Oral Drug Delivery System. *Drug Development and Industrial Pharmacy,* 33**,** 812-823.

MERINO, G., ALVAREZ, A. I., PULIDO, M. M., MOLINA, A. J., SCHINKEL, A. H. & PRIETO, J. G. 2006. Breast cancer resistance protein (BCRP/ABCG2) transports fluoroquinolone antibiotics and affects their oral availability, pharmacokinetics, and milk secretion. *Drug Metabolism and Disposition,* 34**,** 690-695.

METCALFE, P. D. & THOMAS, S. 2010. Challenges in the prediction and modeling of oral absorption and bioavailability. *Current Opinion in Drug Discovery & Development,* 13**,** 104-110.

MICHIELAN, L., TERFLOTH, L., GASTEIGER, J. & MORO, S. 2009. Comparison of Multilabel and Single-Label Classification Applied to the Prediction of the Isoform Specificity of Cytochrome P450 Substrates. *Journal of Chemical Information and Modeling,* 49**,** 2588-2605.

MILLER, J. M., BEIG, A., KRIEG, B. J., CARR, R. A., BORCHARDT, T. B., AMIDON, G. E., AMIDON, G. L. & DAHAN, A. 2011. The Solubility-Permeability Interplay: Mechanistic Modeling and Predictive Application of the Impact of Micellar Solubilization on Intestinal Permeation. *Molecular Pharmaceutics,* 8**,** 1848-1856.

MILLER, J. M., DAHAN, A., GUPTA, D., VARGHESE, S. & AMIDON, G. L. 2010. Enabling the Intestinal Absorption of Highly Polar Antiviral Agents: Ion-Pair Facilitated Membrane Permeation of Zanamivir Heptyl Ester and Guanidino Oseltamivir. *Molecular Pharmaceutics,* 7**,** 1223-1234.

MIN-LING, Z. & ZHI-HUA, Z. 2005. A k-nearest neighbor based algorithm for multi-label classification. *In:* Granular Computing, 2005 IEEE International Conference on, 25-27 July 2005 2005. 718-721 Vol. 2.

MINUESA, G., VOLK, C., MOLINA-ARCAS, M., GORBOULEV, V., ERKIZIA, I., ARNDT, P., CLOTET, B., PASTOR-ANGLADA, M., KOEPSELL, H. & MARTINEZ-PICADO, J. 2009. Transport of Lamivudine (-)-beta-L-2 ',3 '-Dideoxy-3 '-thiacytidine and High-Affinity Interaction of Nucleoside Reverse Transcriptase

Inhibitors with Human Organic Cation Transporters 1, 2, and 3 (vol 329, pg 252, 2009). *Journal of Pharmacology and Experimental Therapeutics,* 329**,** 1187-1187.
MIZUNO, N., NIWA, T., YOTSUMOTO, Y. & SUGIYAMA, Y. 2003. Impact of drug transporter studies on drug discovery and development. *Pharmacological Reviews,* 55**,** 425-461.
MOE 2012. Molecular Operating Environment (MOE), v2012.10. Montreal, QC: Chemical Computing Group Inc.
MOE 2014. QuaSAR-Descriptor help file [Online] Available: http://www.chemcomp.com/journal/descr.htm [Accessed 14 Jan 2014].
MORRIS, M. E., LEE, H. J. & PREDKO, L. M. 2003. Gender differences in the membrane transport of endogenous and exogenous compounds. *Pharmacological Reviews,* 55**,** 229-240.
MORRISSEY, K. M., WEN, C. C., JOHNS, S. J., ZHANG, L., HUANG, S. M. & GIACOMINI, K. M. 2012. The UCSF-FDA TransPortal: a public drug transporter database. *Clinical Pharmacology & Therapeutics,* 92**,** 545-6.
NELSON, T. M. & JURS, P. C. 1994. Prediction of Aqueous Solubility of Organic Compounds. *Journal of Chemical Information and Computer Sciences,* 34**,** 601-609.
NETZEVA, T. I., WORTH, A., ALDENBERG, T., BENIGNI, R., CRONIN, M. T., GRAMATICA, P., JAWORSKA, J. S., KAHN, S., KLOPMAN, G., MARCHANT, C. A., MYATT, G., NIKOLOVA-JELIAZKOVA, N., PATLEWICZ, G. Y., PERKINS, R., ROBERTS, D., SCHULTZ, T., STANTON, D. W., VAN DE SANDT, J. J., TONG, W., VEITH, G. & YANG, C. 2005. Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. The report and recommendations of ECVAM Workshop 52. *ATLA: Alternatives to Lab Animals,* 33**,** 155-173.
NEWBY, D., FREITAS, A. A. & GHAFOURIAN, T. 2013a. Coping with Unbalanced Class Data Sets in Oral Absorption Models. *Journal of Chemical Information and Modeling,* 53**,** 461-474.
NEWBY, D., FREITAS, A. A. & GHAFOURIAN, T. 2013b. Pre-processing Feature Selection for Improved C&RT Models for Oral Absorption. *Journal of Chemical Information and Modeling,* 53**,** 2730-2742.
NEWBY, D., FREITAS, A. A. & GHAFOURIAN, T. 2014a. Comparing Multi-Label Classification Methods for Provisional Biopharmaceutics Class Prediction. *Molecular Pharmaceutics,* (In Press).
NEWBY, D., FREITAS, A. A. & GHAFOURIAN, T. 2014b. Decision trees to characterise the roles of permeability and solubility on the prediction of oral absorption. *European Journal of Medicinal Chemistry,* (Accepted).
NIWA, T. 2003. Using general regression and probabilistic neural networks to predict human intestinal absorption with topological descriptors derived from two-dimensional chemical structures. *Journal of Chemical Information and Computer Sciences,* 43**,** 113-119.
NORDQVIST, A., NILSSON, J., LINDMARK, T., ERIKSSON, A., GARBERG, P. & KIHLÉN, M. 2004. A General Model for Prediction of Caco-2 Cell Permeability. *QSAR & Combinatorial Science,* 23**,** 303-310.
NORINDER, U., OSTERBERG, T. & ARTURSSON, P. 1999. Theoretical calculation and prediction of intestinal absorption of drugs in humans using MolSurf parametrization and PLS statistics. *European Journal of Pharmaceutical Sciences,* 8**,** 49-56.
NOYES, A. A. & WHITNEY, W. R. 1897. The rate of solution of solid substances in their own solutions. *Journal of the American Chemical Society,* 19**,** 930-934.
OECD. 2007. *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models* [Online]. Available:

http://search.oecd.org/officialdocuments/displaydocumentpdf/?doclanguage=en&cote=env/jm/mono(2007)2 [Accessed 3 April 2013].
OPREA, T. I. 2000. Property distribution of drug-related chemical databases. *Journal of Computer-Aided Molecular Design,* 14**,** 251-264.
OPREA, T. I. & MATTER, H. 2004. Integrating virtual screening in lead discovery. *Current Opinion in Chemical Biology,* 8**,** 349-358.
PADE, V. & STAVCHANSKY, S. 1998. Link between drug absorption solubility and permeability measurements in Caco-2 cells. *Journal of Pharmaceutical Sciences,* 87**,** 1604-1607.
PAINE, M. F., HART, H. L., LUDINGTON, S. S., HAINING, R. L., RETTIE, A. E. & ZELDIN, D. C. 2006. The human intestinal cytochrome P450 "pie". *Drug Metabolism and Disposition,* 34**,** 880-886.
PAL, A., BRASSEUR, J. G. & ABRAHAMSSON, B. 2007. A stomach road or "Magenstrasse" for gastric emptying. *Journal of Biomechanics,* 40**,** 1202-1210.
PALM, K., LUTHMAN, K., ROS, J., GRASJO, J. & ARTURSSON, P. 1999. Effect of molecular charge on intestinal epithelial drug transport: pH-dependent transport of cationic drugs. *Journal of Pharmacology and Experimental Therapeutics,* 291**,** 435-443.
PALM, K., LUTHMAN, K., UNGELL, A. L., STRANDLUND, G. & ARTURSSON, P. 1996. Correlation of drug absorption with molecular surface properties. *Journal of Pharmaceutical Sciences,* 85**,** 32-39.
PALM, K., STENBERG, P., LUTHMAN, K. & ARTURSSON, P. 1997. Polar molecular surface properties predict the intestinal absorption of drugs in humans. *Pharmaceutical Research,* 14**,** 568-571.
PANG, K. S. 2003. Modeling of intestinal drug absorption: Roles of transporters and metabolic enzymes (for the Gillette Review Series). *Drug Metabolism and Disposition,* 31**,** 1507-1519.
PARPINELLI, R S., LOPES, H.S., & FREITAS, A.A. "Data mining with an ant colony optimization algorithm." *Evolutionary Computation, IEEE Transactions on* 6.4 (2002): 321-332.
PARSONS, R. L. 1977. Drug Absorption in Gastrointestinal-Disease with Particular Rerefence to Malabsorption-Syndromes. *Clinical Pharmacokinetics,* 2**,** 45-60.
PEREZ, P. A. C., SANZ, M. B., TORRES, L. R., AVALOS, R. C., GONZALEZ, M. P. & DIAZ, H. G. 2004. A topological sub-structural approach for predicting human intestinal absorption of drugs. *European Journal of Medicinal Chemistry,* 39**,** 905-916.
PETRI, N. & LENNERNÄS, H. 2003. *In vivo* Permeability Studies in the Gastrointestinal Tract of Humans. *In:* WATERBEEMD, H. V. D., LENNERNÄS, H. & ARTURSSON, P. (eds.) *Drug Bioavailability Estimation of Solubility, Permeability, Absorption and Bioavailability*. Weinheim: Wiley-VCH.
PHAM-THE, H., GONZALEZ-ALVAREZ, I., BERMEJO, M., SANJUAN, V. M., CENTELLES, I., GARRIGUES, T. M. & CABRERA-PEREZ, M. A. 2011. In Silico Prediction of Caco-2 Cell Permeability by a Classification QSAR Approach. *Molecular Informatics,* 30**,** 376-385.
PHAM-THE, H., GARRIGUES, T., BERMEJO, M., GONZÁLEZ-ÁLVAREZ, I., MONTEAGUDO, M. C. & CABRERA-PÉREZ, M. Á. 2013a. Provisional Classification and in Silico Study of Biopharmaceutical System Based on Caco-2 Cell Permeability and Dose Number. *Molecular Pharmaceutics,* 10**,** 2445-2461.
PHAM-THE, H., GONZÁLEZ-ÁLVAREZ, I., BERMEJO, M., GARRIGUES, T., LE-THI-THU, H. & CABRERA-PÉREZ, M. Á. 2013b. The Use of Rule-Based and QSPR Approaches in ADME Profiling: A Case Study on Caco-2 Permeability. *Molecular Informatics,* 32**,** 459-479.

PINTO, M., ROBINELEON, S., APPAY, M. D., KEDINGER, M., TRIADOU, N., DUSSAULX, E., LACROIX, B., SIMONASSMANN, P., HAFFEN, K., FOGH, J. & ZWEIBAUM, A. 1983. Enterocyte-like differentiation and polarization of the human-colon carcinoma cell-line caco-2 in culture. *Biology of the Cell,* 47**,** 323-330.

PLATTS, J. A., BUTINA, D., ABRAHAM, M. H. & HERSEY, A. 1999. Estimation of Molecular Linear Free Energy Relation Descriptors Using a Group Contribution Approach. *Journal of Chemical Information and Computer Sciences,* 39**,** 835-845.

POCOCK, G. & RICHARDS, C. R. 2009. *The Human Body An Introduction for the Biomedical and Health Sciences,* Oxford, Oxford University Press.

POTTHAST, H., DRESSMAN, J., JUNGINGER, H., MIDHA, K., OESER, H., SHAH, V., VOGELPOEL, H. & BARENDS, D. 2005. Biowaiver monographs for immediate release solid oral dosage forms: Ibuprofen. *Journal of Pharmaceutical Sciences,* 94**,** 2121-2131.

PRESCOTT, L. F. 1974. GASTRIC-EMPTYING AND DRUG ABSORPTION. *British Journal of Clinical Pharmacology,* 1**,** 189-190.

QUAN, Y., JIN, Y., FARIA, T. N., C. A. TILFORD, C. A., HE, A., WALL, D. A., SMITH, R. L. & VIG, B. S. 2012. Expression Profile of Drug and Nutrient Absorption Related Genes in Madin-Darby Canine Kidney (MDCK) Cells Grown under Differentiation Conditions. *Pharmaceutics,* 4**,** 314-333.

QUECKENBERG, C. & FUHR, U. 2009. Influence of posture on pharmacokinetics. *European Journal of Clinical Pharmacology,* 65**,** 109-119.

QUINLAN, J. R. 1979. *Discovering rules from large collections of examples: A case study,* Edinburgh, Edinburgh University Press.

QUINLAN, J. R. 1993. *C4.5: programs for machine learning,* San Francisco, Morgan Kaufmann Publishers Inc.

RAEVSKY, O. A., GRIGOR'EV, V. Y., POLIANCZYK, D. E., RAEVSKAJA, O. E. & DEARDEN, J. C. 2014. Calculation of Aqueous Solubility of Crystalline Un-Ionized Organic Chemicals and Drugs Based on Structural Similarity and Physicochemical Descriptors. *Journal of Chemical Information and Modeling,* 54**,** 683-691.

RAUB, T. J. 2005. P-Glycoprotein Recognition of Substrates and Circumvention through Rational Drug Design. *Molecular Pharmaceutics,* 3**,** 3-25.

READ, J., PFAHRINGER, B., HOLMES, G. & FRANK, E. 2011. Classifier chains for multi-label classification. *Machine Learning,* 85**,** 333-359.

REYNOLDS, D. P., LANEVSKIJ, K., JAPERTAS, P., DIDZIAPETRIS, R. & PETRAUSKAS, A. 2009. Ionization-Specific Analysis of Human Intestinal Absorption. *Journal of Pharmaceutical Sciences,* 98**,** 4039-4054.

RICHENS, A. 1975. Drug interactions and lethal drug combinations. *Journal of Clinical Pathology,* 28**,** 94-98.

ROSENBAUM, S. E. 2011. *Basic Pharmacokinetics and Pharmacodynamics: An Integrated Textbook and Computer Simulations,* New Jersey, Wiley.

ROSS, D. L. & RILEY, C. M. 1990. Aqueous solubilities of some variously substituted quinolone antimicrobials. *International Journal of Pharmaceutics,* 63**,** 237-250.

RUEMMELE, F. M., SEIDMAN, E. G. & LENTZE, M. J. 2002. Regulation of intestinal epithelial cell apoptosis and the pathogenesis of inflammatory bowel disorders. *Journal of Pediatric Gastroenterology and Nutrition,* 34**,** 254-260.

RYDZEWSKI, M. R. 2008. *Real World Drug Discovery A Chemist's Guide to Biotech and Pharmaceutical Research,* Oxford, Elsevier.

SAEYS, Y., INZA, I. & LARRANAGA, P. 2007. A review of feature selection techniques in bioinformatics. *Bioinformatics,* 23**,** 2507-2517.

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**,** 4791-4810.

SAI, Y. & TSUJI, A. 2004. Transporter-mediated drug delivery: recent progress and experimental approaches. *Drug Discovery Today,* 9**,** 712-20.
SALAHINEJAD, M., LE, T. C. & WINKLER, D. A. 2013. Aqueous Solubility Prediction: Do Crystal Lattice Interactions Help? *Molecular Pharmaceutics,* 10**,** 2757-2766.
SANGHVI, T., NI, N. & YALKOWSKY, S. H. 2001. A simple modified absorption potential. *Pharmaceutical Research,* 18**,** 1794-1796.
SAVJANI, K. T., GAJJAR, A. K. & SAVJANI, J. K. 2012. Drug solubility: importance and enhancement techniques. *ISRN pharmaceutics,* 2012**,** 195727-195727.
SCHAPIRE, R. & SINGER, Y. 2000. BoosTexter: A Boosting-based System for Text Categorization. *Machine Learning,* 39**,** 135-168.
SCHIETGAT, L., VENS, C., STRUYF, J., BLOCKEEL, H., KOCEV, D. & DZEROSKI, S. 2010. Predicting gene function using hierarchical multi-label decision tree ensembles. *BMC Bioinformatics,* 11.
SCHWARTZ, J. B. 2003. The influence of sex on pharmacokinetics. *Clinical Pharmacokinetics,* 42**,** 107-121.
SECKER, A., DAVIES, M. N., FREITAS, A. A., CLARK, E. B., TIMMIS, J. & FLOWER, D. R. 2010. Hierarchical classification of G-protein-coupled receptors with data-driven selection of attributes and classifiers. *International Journal of Data Mining and Bioinformatics,* 4**,** 191-210.
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.
SERAJUDDIN, A. T. M., RANADIVE, S. A. & MAHONEY, E. M. 1991. Relative Lipophilicities, Solubilities, And Structure Pharmacological Considerations Of 3-Hydroxy-3-Methylglutaryl-Coenzyme-A (Hmg-Coa) Reductase Inhibitors Pravastatin, Lovastatin, Mevastatin, And Simvastatin. *Journal of Pharmaceutical Sciences,* 80**,** 830-834.
SHAH, P., JOGANI, V., BAGCHI, T. & MISRA, A. 2006. Role of Caco-2 cell monolayers in prediction of intestinal drug absorption. *Biotechnology Progress,* 22**,** 186-198.
SHAH, S. C. & KUSIAK, A. 2004. Data mining and genetic algorithm based gene/SNP selection. *Artificial Intelligence in Medicine,* 31**,** 183-196.
SHAO, H., LI, G., LIU, G. & WANG, Y. 2013. Symptom selection for multi-label data of inquiry diagnosis in traditional Chinese medicine. *Science China Information Sciences,* 56**,** 1-13.
SHEN, J., CHENG, F. X., XU, Y., LI, W. H. & TANG, Y. 2010. Estimation of ADME Properties with Substructure Pattern Recognition. *Journal of Chemical Information and Modeling,* 50**,** 1034-1041.
SHEN, L. 2009. Functional morphology of the gastrointestinal tract. *Current Topics in Microbiology and Immunology,* 337**,** 1-35.
SHERER, E. C., VERRAS, A., MADEIRA, M., HAGMANN, W. K., SHERIDAN, R. P., ROBERTS, D., BLEASBY, K. & CORNELL, W. D. 2012. QSAR Prediction of Passive Permeability in the LLC-PK1 Cell Line: Trends in Molecular Properties and Cross-Prediction of Caco-2 Permeabilities. *Molecular Informatics,* 31**,** 231-245.
SHUGARTS, S. & BENET, L. Z. 2009. The Role of Transporters in the Pharmacokinetics of Orally Administered Drugs. *Pharmaceutical Research,* 26**,** 2039-2054.
SILLA, C., JR. & FREITAS, A. 2011. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery,* 22**,** 31-72.
SILVERTHORN, D. U. 2001. *Human Physiology An Integrated Approach,* New Jersey, Prentice Hall.
SINGER, S. J. & NICHOLSON, G. L. 1972. The fluid mosaic model of the structure of cell membranes. *Science,* 175**,** 720-731.

SINKO, P. J. 1999. Drug selection in early drug development: screening for acceptable pharmacokinetic properties using combined in vitro and computational approaches. *Current Opinion in Drug Discovery & Development,* 2**,** 42-8.

SMITH, D., ARTURSSON, P., AVDEEF, A., DI, L., ECKER, G. F., FALLER, B., HOUSTON, J. B., KANSY, M., KERNS, E. H., KRÄMER, S. D., LENNERNÄS, H., VAN DE WATERBEEMD, H., SUGANO, K. & TESTA, B. 2014. Passive Lipoidal Diffusion and Carrier-Mediated Cell Uptake Are Both Important Mechanisms of Membrane Permeation in Drug Disposition. *Molecular Pharmaceutics,* 11**,** 1727-1738.

SPECIAN, R. D. & OLIVER, M. G. 1991. Functional Biology Of Intestinal Goblet Cells. *American Journal of Physiology,* 260**,** C183-C193.

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. *Electronic Notes in Theoretical Computer Science,* 292**,** 135-151.

STEGEMANN, S., LEVEILLER, F., FRANCHI, D., DE JONG, H. & LINDEN, H. 2007. When poor solubility becomes an issue: From early stage to proof of concept. *European Journal of Pharmaceutical Sciences,* 31**,** 249-261.

STENBERG, P., LUTHMAN, K. & ARTURSSON, P. 2000. Virtual screening of intestinal drug permeability. *Journal of Controlled Release,* 65**,** 231-243.

STENBERG, P., NORINDER, U., LUTHMAN, K. & ARTURSSON, P. 2001. Experimental and computational screening models for the prediction of intestinal drug absorption. *Journal of Medicinal Chemistry,* 44**,** 1927-1937.

STOUCH, T. R., KENYON, J. R., JOHNSON, S. R., CHEN, X. Q., DOWEYKO, A. & LI, Y. 2003. In silico ADME/Tox: why models fail. *Journal of Computer-Aided Molecular Design,* 17**,** 83-92.

SUENDERHAUF, C., HAMMANN, F., MAUNZ, A., HELMA, C. & HUWYLER, J. 2011. Combinatorial QSAR Modeling of Human Intestinal Absorption. *Molecular Pharmaceutics,* 8**,** 213-224.

SUGANO, K. 2011. Fraction of a dose absorbed estimation for structurally diverse low solubility compounds. *International Journal of Pharmaceutics,* 405**,** 79-89.

SUGANO, K., KANSY, M., ARTURSSON, P., AVDEEF, A., BENDELS, S., DI, L., ECKER, G. F., FALLER, B., FISCHER, H., GEREBTZOFF, G., LENNERNAES, H. & SENNER, F. 2010. Coexistence of passive and carrier-mediated processes in drug transport. *Nature Reviews Drug Discovery,* 9**,** 597-614.

SUSANTO PARK, M. & CHANG, J., H. 2011. Absorption of drugs via passive diffusion and carrier-mediated pathways. *In:* XIAOLING, L. & MING, H. (eds.) *Oral Bioavailability: Basic Principles, Advanced Concepts, and Applications*. 1 ed. New Jersey: Wiley.

TAFT, W., R. 1952. Polar and steric substituent constants for aliphatic and o- benzoate groups from rates of esterification and hydrolysis of esters. *Journal of the American Chemical Society,* 74**,** 3120-3128.

TAKAGI, T., RAMACHANDRAN, C., BERMEJO, M., YAMASHITA, S., YU, L. X. & AMIDON, G. L. 2006. A Provisional Biopharmaceutical Classification of the Top 200 Oral Drug Products in the United States, Great Britain, Spain, and Japan. *Molecular Pharmaceutics,* 3**,** 631-643.

TALEVI, A., GOODARZI, M., ORTIZ, E. V., DUCHOWICZ, P. R., BELLERA, C. L., PESCE, G., CASTRO, E. A. & BRUNO-BLANCH, L. E. 2011. Prediction of drug intestinal absorption by new linear and non-linear QSPR. *European Journal of Medicinal Chemistry,* 46**,** 218-228.

TAN, P. N., STEINBACH, M. & KUMAR, V. 2006. *Introduction to Data Mining,* Boston, Pearson International Edition.

TAVELIN, S., TAIPALENSUU, J., SODERBERG, L., MORRISON, R., CHONG, S. H. & ARTURSSON, P. 2003. Prediction of the oral absorption of low-permeability drugs

using small intestine-like 2/4/A1 cell monolayers. *Pharmaceutical Research,* 20**,** 397-405.

TESTA, B., CRIVORI, P., REIST, M. & CARRUPT, P. A. 2000. The influence of lipophilicity on the pharmacokinetic behavior of drugs: Concepts and examples. *Perspectives in Drug Discovery and Design,* 19**,** 179-211.

THELEN, K. & DRESSMAN, J. B. 2009. Cytochrome P450-mediated metabolism in the human gut wall. *Journal of Pharmacy and Pharmacology,* 61**,** 541-558.

THOMAS, V. H., BHATTACHAR, S., HITCHINGHAM, L., ZOCHARSKI, P., NAATH, M., SURENDRAN, N., STONER, C. L. & EL-KATTAN, A. 2006. The road map to oral bioavailability: an industrial perspective. *Expert Opinion on Drug Metabolism & Toxicology,* 2**,** 591-608.

TIAN, S., LI, Y., WANG, J., ZHANG, J. & HOU, T. 2011. ADME Evaluation in Drug Discovery. 9. Prediction of Oral Bioavailability in Humans Based on Molecular Properties and Structural Fingerprints. *Molecular Pharmaceutics,* 8**,** 841-851.

TITUS, R., KASTENMEIER, A. & OTTERSON, M. F. 2013. Consequences of Gastrointestinal Surgery on Drug Absorption. *Nutrition in Clinical Practice,* 28**,** 429-436.

TODESCHINI, R. & CONSONNI, V. 2000. *Handbook of Molecular Descriptors,* Verlag, Wiley-VCH.

TROPSHA, A. 2010. Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular Informatics,* 29**,** 476-488.

TROPSHA, A., GRAMATICA, P. & GOMBAR, V. K. 2003. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR & Combinatorial Science,* 22**,** 69-77.

TSOUMAKAS, G. & KATAKIS, I. 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM),* 3**,** 1-13.

TURNER, J. V., GLASS, B. D. & AGATONOVIC-KUSTRIN, S. 2003. Prediction of drug bioavailability based on molecular structure. *Analytica Chimica Acta,* 485**,** 89-102.

UNGELL, A.-L. B. 2004. Caco-2 replace or refine? *Drug Discovery Today: Technologies,* 1**,** 423-430.

UNGELL, A. L., NYLANDER, S., BERGSTRAND, S., SJOBERG, A. & LENNERNAS, H. 1998. Membrane transport of drugs in different regions of the intestinal tract of the rat. *Journal of Pharmaceutical Sciences,* 87**,** 360-366.

VAN BREEMEN, R. B. & LI, L. 2005. Caco-2 cell permeability assays to measure drug absorption. *Expert Opinion on Drug Metabolism & Toxicology,* 1**,** 175-185

VAN DE WATERBEEMD, H. & GIFFORD, E. 2003. ADMET in silico modelling: towards prediction paradise? *Nature Reviews Drug Discovery,* 2**,** 192-204.

VAN DE WATERBEEMD, H., SMITH, D. A. & JONES, B. C. 2001. Lipophilicity in PK design: methyl, ethyl, futile. *Journal of Computer-Aided Molecular Design,* 15**,** 273-286.

VAN GELDER, J., WITVROUW, M., PANNECOUQUE, C., HENSON, G., BRIDGER, G., NAESENS, L., DE CLERCQ, E., ANNAERT, P., SHAFIEE, M., VAN DEN MOOTER, G., KINGET, R. & AUGUSTIJNS, P. 1999. Evaluation of the potential of ion pair formation to improve the oral absorption of two potent antiviral compounds, AMD3100 and PMPA. *International Journal of Pharmaceutics,* 186**,** 127-136.

VARMA, M. V., GARDNER, I., STEYN, S. J., NKANSAH, P., ROTTER, C. J., WHITNEY-PICKETT, C., ZHANG, H., DI, L., CRAM, M., FENNER, K. S. & EL-KATTAN, A. F. 2012. pH-Dependent Solubility and Permeability Criteria for Provisional Biopharmaceutics Classification (BCS and BDDCS) in Early Drug Discovery. *Molecular Pharmaceutics,* 9**,** 1199-1212.

VARMA, M. V. S., OBACH, R. S., ROTTER, C., MILLER, H. R., CHANG, G., STEYN, S. J., EL-KATTAN, A. & TROUTMAN, M. D. 2010. Physicochemical Space for

Optimum Oral Bioavailability: Contribution of Human Intestinal Absorption and First-Pass Elimination. *Journal of Medicinal Chemistry,* 53**,** 1098-1108.

VARMA, M. V. S., SATEESH, K. & PANCHAGNULA, R. 2005. Functional role of P-glycoprotein in limiting intestinal absorption of drugs: Contribution of passive permeability to P-glycoprotein mediated efflux transport. *Molecular Pharmaceutics,* 2**,** 12-21.

VEBER, D. F., JOHNSON, S. R., CHENG, H.-Y., SMITH, B. R., WARD, K. W. & KOPPLE, K. D. 2002. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *Journal of Medicinal Chemistry,* 45**,** 2615-2623.

VOLPE, D. A. 2008. Variability in Caco-2 and MDCK cell-based intestinal permeability assays. *Journal of Pharmaceutical Sciences,* 97**,** 712-725.

WANCHANA, S., YAMASHITA, F., HARA, H., FUJIWARA, S. I., AKAMATSU, M. & HASHIDA, M. 2004. Two- and three-dimensional QSAR of carrier-mediated transport of beta-lactam antibiotics in Caco-2 cells. *Journal of Pharmaceutical Sciences,* 93**,** 3057-3065.

WANG, E. J., CASCIANO, C. N., CLEMENT, R. P. & JOHNSON, W. W. 2001. HMG-CoA reductase inhibitors (statins) characterized as direct inhibitors of P-glycoprotein. *Pharmaceutical Research,* 18**,** 800-806.

WANG, J. M. & HOU, T. J. 2011. Recent Advances on Aqueous Solubility Prediction. *Combinatorial Chemistry & High Throughput Screening,* 14**,** 328-338.

WANG, J. M., XIE, X. Q., HOU, T. J. & XU, X. J. 2007. Fast approaches for molecular polarizability calculations. *Journal of Physical Chemistry A,* 111**,** 4443-4448.

WANG, Y., LI, Y., DING, J., JIANG, Z. & CHANG, Y. 2008. Estimation of bioconcentration factors using molecular electro-topological state and flexibility. *SAR and QSAR in Environmental Research,* 19**,** 375-395.

WASSVIK, C. M., HOLMÉN, A. G., BERGSTRÖM, C. A. S., ZAMORA, I. & ARTURSSON, P. 2006. Contribution of solid-state properties to the aqueous solubility of drugs. *European Journal of Pharmaceutical Sciences,* 29**,** 294-305.

WASSVIK, C. M., HOLMEN, A. G., DRAHEIM, R., ARTURSSON, P. & BERGSTROM, C. A. S. 2008. Molecular characteristics for solid-state limited solubility. *Journal of Medicinal Chemistry,* 51**,** 3035-3039.

WEGNER, J. K., FROHLICH, H. & ZELL, A. 2004. Feature selection for descriptor based classification models. 2. Human intestinal absorption (HIA). *Journal of Chemical Information and Computer Sciences,* 44**,** 931-939.

WELLING, P. G. 1984. INTERACTIONS AFFECTING DRUG ABSORPTION. *Clinical Pharmacokinetics,* 9**,** 404-434.

WESSEL, M. D., JURS, P. C., TOLAN, J. W. & MUSKAL, S. M. 1998. Prediction of human intestinal absorption of drug compounds from molecular structure. *Journal of Chemical Information and Computer Sciences,* 38**,** 726-735.

WHITE, R. E. 2000. High-throughput screening in drug metabolism and pharmacokinetic support of drug discovery. *Annual Review of Pharmacology and Toxicology,* 40**,** 133-157.

WILLIAMS, H. D., TREVASKIS, N. L., CHARMAN, S. A., SHANKER, R. M., CHARMAN, W. N., POUTON, C. W. & PORTER, C. J. 2013. Strategies to address low drug solubility in discovery and development. *Pharmacological Reviews,* 65**,** 315-499.

WINIWARTER, S., BONHAM, N. M., AX, F., HALLBERG, A., LENNERNÄS, H. & KARLÉN, A. 1998. Correlation of Human Jejunal Permeability (in Vivo) of Drugs with Experimentally and Theoretically Derived Parameters. A Multivariate Data Analysis Approach. *Journal of Medicinal Chemistry,* 41**,** 4939-4949.

WONG, W. W. L. & BURKOWSKI, F. J. 2011. Using Kernel Alignment to Select Features of Molecular Descriptors in a QSAR Study. *IEEE Transactions on Computational Biology and Bioinformatics,* 8**,** 1373-1384.

WU, C. Y. & BENET, L. Z. 2005. Predicting drug disposition via application of BCS: Transport/absorption/elimination interplay and development of a biopharmaceutics drug disposition classification system. *Pharmaceutical Research,* 22**,** 11-23.

XU, L. & ZHANG, W.-J. 2001. Comparison of different methods for variable selection. *Analytica Chimica Acta,* 446**,** 475-481.

XUE, Y., LI, Z. R., YAP, C. W., SUN, L. Z., CHEN, X. & CHEN, Y. Z. 2004. Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents. *Journal of Chemical Information and Computer Sciences,* 44**,** 1630-1638.

YALKOWSKY, S. H., JOHNSON, J. L. H., SANGHVI, T. & MACHATHA, S. G. 2006. A 'rule of unity' for human intestinal absorption. *Pharmaceutical Research,* 23**,** 2475-2481.

YAN, A., WANG, Z. & CAI, Z. 2008. Prediction of Human Intestinal Absorption by GA Feature Selection and Support Vector Machine Regression. *International Journal of Molecular Sciences,* 9**,** 1961-1976.

YANG, Y. 1999. An evaluation of statistical approaches to text categorization. *Information Retrieval,* 1**,** 69-90.

YANG, Y., ENGKVIST, O., LLINAS, A. & CHEN, H. 2012. Beyond size, ionization state, and lipophilicity: influence of molecular topology on absorption, distribution, metabolism, excretion, and toxicity for druglike compounds. *Journal of Medicinal Chemistry,* 55**,** 3667-77.

YEE, S. Y. 1997. In vitro permeability across Caco3 cells (colonic) can predict in vivo (small intestinal) absorption in man - Fact or myth. *Pharmaceutical Research,* 14**,** 763-766.

YOSHIDA, K., MAEDA, K. & SUGIYAMA, Y. 2013. Hepatic and Intestinal Drug Transporters: Prediction of Pharmacokinetic Effects Caused by Drug-Drug Interactions and Genetic Polymorphisms. *Annual Review of Pharmacology and Toxicology,* 53**,** 581-612.

YOUDEN, W. J. 1950. Index for rating diagnostic tests. *Cancer,* 3**,** 32-35.

YOUNG, D., MARTIN, T., VENKATAPATHY, R. & HARTEN, P. 2008. Are the Chemical Structures in Your QSAR Correct? *QSAR & Combinatorial Science,* 27**,** 1337-1345.

YU, H. S. & ADEDOYIN, A. 2003. ADME-Tox in drug discovery: integration of experimental and computational technologies. *Drug Discovery Today,* 8**,** 852-861.

YUEN, K. H. 2010. The transit of dosage forms through the small intestine. *International Journal of Pharmaceutics,* 395**,** 9-16.

ZAKERI-MILANI, P., TAJERZADEH, H., ISLAMBOLCHILAR, Z., BARZEGAR, S. & VALIZADEH, H. 2006. The relation between molecular properties of drugs and their transport across the intestinal membrane. *DARU Journal of Pharmaceutical Sciences,* 14**,** 164-171.

ZANG, Q., ROTROFF, D. M. & JUDSON, R. S. 2013. Binary Classification of a Large Collection of Environmental Chemicals from Estrogen Receptor Assays by Quantitative Structure–Activity Relationship and Machine Learning Methods. *Journal of Chemical Information and Modeling,* 53**,** 3244-3261.

ZHANG, J. H., CHUNG, T. D. Y. & OLDENBURG, K. R. 2000. Confirmation of primary active substances from high throughput screening of chemical and biological populations: A statistical approach and practical considerations. *Journal of Combinatorial Chemistry,* 2**,** 258-265.

ZHANG, Y. C. & BENET, L. Z. 2001. The gut as a barrier to drug absorption - Combined role of cytochrome P450 3A and P-glycoprotein. *Clinical Pharmacokinetics,* 40**,** 159-168.

ZHAO, Y. H., ABRAHAM, M. H., LE, J., HERSEY, A., LUSCOMBE, C. N., BECK, G., SHERBORNE, B. & COOPER, I. 2002. Rate-limited steps of human oral absorption and QSAR studies. *Pharmaceutical Research,* 19**,** 1446-1457.

ZHAO, Y. H., LE, J., ABRAHAM, M. H., HERSEY, A., EDDERSHAW, P. J., LUSCOMBE, C. N., BOUTINA, D., BECK, G., SHERBORNE, B., COOPER, I. & PLATTS, J. A. 2001. Evaluation of human intestinal absorption data and subsequent derivation of a quantitative structure-activity relationship (QSAR) with the Abraham descriptors. *Journal of Pharmaceutical Sciences,* 90**,** 749-784.

ZHU, J. Y., WANG, J. M., YU, H. D., LI, Y. Y. & HOU, T. J. 2011. Recent Developments of In Silico Predictions of Oral Bioavailability. *Combinatorial Chemistry & High Throughput Screening,* 14**,** 362-374.

ZHU, S., JI, X., XU, W. & GONG, Y. 2005. Multi-labelled classification using maximum entropy method. *In:* Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, 2005. ACM, 274-281.

Appendices

The supporting information directly mentioned in this thesis can be found in the appendices below. The supporting information for each chapter has been grouped into a single appendix where possible. Supporting information too large and not directly mentioned in the main text can be found on the accompanying disk.

Appendix 1: Supporting Information for Chapters 6 and 7

Table A1. 1. Summary of molecular descriptors sets used in chapter 6 (Descriptor sets 1-5) and molecular descriptors sets (Descriptor sets 1-4) in chapter 7

Descriptor	Model/ Set	Description
aliphatic rings(5)	2	Number of 5 aliphatic rings
aromatic rings(6)	5	Number of 6 aromatic rings
ACD_LogP	4	Octanol/water partition coefficient calculated by ACD
ACDLogD2	3	Apparent Distribution coefficient at PH 2 calculated by ACD
ACDLogD5.5	1,3,5	Apparent Distribution coefficient at pH5.5 calculated by ACD
ACDLogD7.4	2	Apparent Distribution coefficient at pH7.4 calculated by ACD
ACD_Density	1	Mass per unit volume of a molecule calculated by dividing MW by MV calculated by ACD
FiAB1	5	Fraction of drugs ionised as anions
HBA	4	The total number of hydrogen bond acceptors of the whole molecule
HBD	4	The total number of hydrogen bonds donors of the whole molecule
Inertia moment 2 size	1	An estimate of an object resistance to changes in its rotation rate
Ka3	2	Kappa alpha 3: atom count which quantifies the extent the heteroatom differs from the reference atom(carbon sp3)
Mass	4	The total mass of the whole molecule
NRo5	3	Number of violations of the rule of five
PSA	2,3	Polar surface Area
RB	4	The total number of rotatable bonds of the whole molecule
SdsssP	3	Sum of atom-type E-state for phosphorous atoms with 3 single and one double bond
SdsssP_acnt	5	Counts of atom-type E-state for phosphorus atoms with 3 single and one double bond.
SHBint2	1,2	Sum of E-state descriptors for potent hydrogen bonds of path length 2
SHBint2_Acnt	2	Counts of internal hydrogen bonds with 3 skeletal bonds between donor and acceptor
SHBint3	2,5	Sum of E-state descriptors for potent hydrogen bonds of path length 3
SHBint7	1,5	Sum of E-state descriptors for potent hydrogen bonds of path length 7
SHBint9	3	Sum of E-state descriptors for potent H2 bonds of path length 9
SHHBd	1,5	Sum of the hydrogen atom level E-state values for all hydrogen atoms bonded to donating atoms.
SpcPolarizability	1,2	Molecular polarizability calculated on the basis of the addictive approach
SsCH3	1	Sum of all (-CH3 -) E-state values in molecule

Appendix 2: Supporting Information for Chapter 8

Table A2. 1. Top 20 molecular descriptors picked by feature selection methods, IGR, CS, RF and RF (MC) used in chapter 8

Rank Number	IGR	CS	RF	RF (MC)
1	N+	ACDLogD7.4	ACDLogD10	VAMP HOMO
2	SHssNH	ACDLogD6.5	ACDLogD6.5	ACD_PSA
3	SssNH	SHHBd	O Atoms	ACDLogD5.5
4	Methyl	ACD_PSA	numHBd	VAMP Heat of Formation
5	SHsNH2	ACDLogD5.5	ACDLogD7.4	SsOH_acnt
6	SdsssP	ACD_LogP	VAMP Mean Polarizability	xp6
7	5-aliphatic rings	ACDLogD10	H-bond Donors	ABSQon
8	Amino	H-bond Acceptors	ACD_PSA	phia
9	ACDLogD10	O Atoms	ACDLogD5.5	ACDLogD7.4
10	VAMP totl Energy	H-bond Donors	SHBint9_Acnt	FU7.4
11	ACDLogD7.4	numHBd	ka1	Heteroatoms
12	VAMP Electronic Energy	numHBa	numHBa	ka1
13	SdsssP_acnt	SHsOH	VAMP Heat of Formation	k0
14	Atoms	SHHBa	SsOH_acnt	SsssCH
15	SHHBd	ABSQ	ACD_FRB	MaxQp
16	Sum of E-State indices	Heteroatoms	LogP	O Atoms
17	SssCH2	SsOH_acnt	Sum of E-State indices	FiA7.4
18	Mass	LogP	Hmin	VAMP Mean Polarizability
19	Randic Topological index	ACDLogD2	VAMP LUMO	SHother
20	Wiener Topological index	SHBint3_Acnt	H-bond Acceptors	numHBa

IGR: Information Gains Ratio; CS: Chi Square; RF: Variable importance using random forest; RF (MC): Variable importance using random forest with higher misclassification costs applied to reduce false positives

Table A2. 2. Top molecular descriptors picked by feature selection methods GRD and GEN used in chapter 8

GEN			GRD
Inertia Moment 2 Size	xvch6	ACDLogD10	N+ FINAL
Randic Topological index	ncirc	Cosmic Total Energy	Mass
Balaban Topological index	knotp	xvch9	O Atoms
Sum of E-State indices	numHBa	xp3	SdsssP
Heteroatoms	numHBd	xp4	Dipole
C Atoms	SHHBd	xp6	SHHBd
N Atoms	Qv	xc4	SHBint3
5-aliphatic rings	k1	xch6	SHBint9_Acnt
6-aromatic rings	ka1	SdCH2	SHssNH
4-rings	ka3	SaaCH	FU7.4
aliphatics	SHBint2	SsssCH	VAMP totl Energy
aromatics	SHBint3	SssNH	VAMP Heat of Formation
Atoms	SHBint4	FiA13	VAMP LUMO
SdCH2_acnt	SHBint5_Acnt	FU13	VAMP HOMO
SsOH_acnt	SHsOH	Shape Flexibility index	ACD_PSA_1
SssO_acnt	SHsNH2	VAMP Heat of Formation	ACD_LogP_1
SdsssP	SHother	VAMP LUMO	ACDLogD5.5
ABSQ	Gmin	VAMP HOMO	ACDLogD6.5
Dipole	xv2	ACD_FRB	ACDLogD7.4
Surface	xvp4	ACD_LogP	ACDLogD10
Volume	xvch3		Cosmic Total Energy
ACDLogD7.4	ACDLogD5.5		

Gen: Genetic Search; GRD: Greedy stepwise

Appendix 3: Supporting Information for Chapter 9

Table A3. 1. Compound outliers as highlighted from Figure 9.3 in chapter 9

Outlier type	Name	Therapeutic indication/Class	Transport Route	Comments	References
Higher permeability in caco-2 compared to MDCK	Oseltamivir	Antiviral	Influx, PEPT1 & Efflux, P-gp	PEPT1 is not detected in MDCK cells; therefore higher permeability for the caco-2 cell lines	(Ogihara *et al*., 2009)
	Loperamide	Opioid/gastrointestinal	Efflux, P-gp	In this case the MDCK permeability is lower compared to caco-2 permeability due to the strain of MDCK used which over expresses the p-gp transporter therefore would expect lower permeability in MDCK-MDR1 cells compared to caco-2	(Varma *et al*., 2005, Troutman and Thakker, 2003)
	Amitriptyline	Antidepressant	Efflux, P-gp	In this case the MDCK permeability is lower compared to caco-2 permeability due to the strain of MDCK used which over expresses the p-gp transporter therefore would expect lower permeability in MDCK-MDR1 cells compared to caco-2	(Faassen *et al*., 2003, Troutman and Thakker, 2003)
	Phenazopyridine	Analgesic	Passive Transcellular	Poor solubility (dissolution limiting) - BCS class 2 compound	(Gao, 2012)
Higher permeability in MDCK compared to caco-2	Sotalol	Antiarrhythmic	Paracellular & influx transporter OATP-A	OATP-A is expressed in caco-2 but not determined in MDCK II. In addition MDCK II cells are leakier compared with caco-2 hence both reasons could contribute to higher absorption through the MDCK cell line	(Liu *et al*., 2012)
	Dicloxacillin	Antibiotic	Influx, PEPT1 & Efflux, P-gp	Higher expression of p-gp in caco-2 and higher efflux of dicloxacillin by this transporter than uptake by PEPT1 could explain higher permeability in MDCK cell line.	(Luckner and Brandsch, 2005, Susanto and Benet, 2002)
	Levodopa	Psychoactive	Influx, LNAA	Higher abundance of LNAA in MDCK cells compared with caco-2, in addition the possibility of paracellular absorption has been proposed and based on the MDCK cell line which is leakier both could explain the higher permeability in MDCK compared to caco-2	(Putnam *et al*., 2002, Lennernas *et al*., 1993)
	Sildenafil	Erectile dysfunction and pulmonary arterial hypertension	Efflux, P-gp & BCRP	BCRP transporter was not determined in MDCK II cells indicating that the combined efflux effect of p-gp and BCRP transporters in caco-2 reduce the permeability more compared to the single efflux transporter in the MDCK cell line hence the higher permeability in the MDCK cell line	(Choi and Song, 2012, Di *et al*., 2011)
	Glipizide	Anti-diabetic	Passive Transcellular	Poor solubility (dissolution limiting) - BCS class 2 compound	(Mehramizi *et al*., 2007)

.

References for Table A3.1

CHOI, M. K. & SONG, I. S. 2012. Characterization of efflux transport of the PDE5 inhibitors, vardenafil and sildenafil. *Journal of Pharmacy and Pharmacology,* 64**,** 1074-1083.

DI, L., WHITNEY-PICKETT, C., UMLAND, J. P., ZHANG, H., ZHANG, X., GEBHARD, D. F., LAI, Y. R., FEDERICO, J. J., DAVIDSON, R. E., SMITH, R., REYNER, E. L., LEE, C., FENG, B., ROTTER, C., VARMA, M. V., KEMPSHALL, S., FENNER, K., EL-KATTAN, A. F., LISTON, T. E. & TROUTMAN, M. D. 2011. Development of a New Permeability Assay Using Low-Efflux MDCKII Cells. *Journal of Pharmaceutical Sciences,* 100**,** 4974-4985.

FAASSEN, F., VOGEL, G., SPANINGS, H. & VROMANS, H. 2003. Caco-2 permeability, P-glycoprotein transport ratios and brain penetration of heterocyclic drugs. *International Journal of Pharmaceutics,* 263**,** 113-122.

GAO, Z. 2012. Development of a Continuous Dissolution/Absorption System—a Technical Note. *AAPS PharmSciTech,* 13**,** 1287-1292.

LENNERNAS, H., NILSSON, D., AQUILONIUS, S. M., AHRENSTEDT, O., KNUTSON, L. & PAALZOW, L. K. 1993. THE EFFECT OF L-LEUCINE ON THE ABSORPTION OF LEVODOPA, STUDIED BY REGIONAL JEJUNAL PERFUSION IN MAN. *British Journal of Clinical Pharmacology,* 35**,** 243-250.

LIU, W., OKOCHI, H., BENET, L. Z. & ZHAI, S. D. 2012. Sotalol Permeability in Cultured-Cell, Rat Intestine, and PAMPA System. *Pharmaceutical Research,* 29**,** 1768-1774.

LUCKNER, P. & BRANDSCH, M. 2005. Interaction of 31 β-lactam antibiotics with the H+/peptide symporter PEPT2: analysis of affinity constants and comparison with PEPT1. *European Journal of Pharmaceutics and Biopharmaceutics,* 59**,** 17-24.

MEHRAMIZI, A., ALIJANI, B., POURFARZIB, M., DORKOOSH, F. A. & RAFIEE – TEHRANI, M. 2007. Solid Carriers for Improved Solubility of Glipizide in Osmotically Controlled Oral Drug Delivery System. *Drug Development and Industrial Pharmacy,* 33**,** 812-823.

OGIHARA, T., KANO, T., WAGATSUMA, T., WADA, S., YABUUCHI, H., ENOMOTO, S., MORIMOTO, K., SHIRASAKA, Y., KOBAYASHI, S. & TAMAI, I. 2009. Oseltamivir (Tamiflu) Is a Substrate of Peptide Transporter 1. *Drug Metabolism and Disposition,* 37**,** 1676-1681.

PUTNAM, W. S., RAMANATHAN, S., PAN, L., TAKAHASHI, L. H. & BENET, L. Z. 2002. Functional characterization of monocarboxylic acid, large neutral amino acid, bile acid and peptide transporters, and P-glycoprotein in MDCK and Caco-2 cells. *Journal of Pharmaceutical Sciences,* 91**,** 2622-2635.

SUSANTO, M. & BENET, L. Z. 2002. Can the enhanced renal clearance of antibiotics in cystic fibrosis patients be explained by P-glycoprotein transport? *Pharmaceutical Research,* 19**,** 457-62.

TROUTMAN, M. D. & THAKKER, D. R. 2003. Novel experimental parameters to quantify the modulation of absorptive and secretory transport of compounds by P-glycoprotein in cell culture models of intestinal epithelium. *Pharmaceutical Research,* 20**,** 1210-1224.

VARMA, M. V. S., SATEESH, K. & PANCHAGNULA, R. 2005. Functional role of P-glycoprotein in limiting intestinal absorption of drugs: Contribution of passive permeability to P-glycoprotein mediated efflux transport. *Molecular Pharmaceutics,* 2**,** 12-21

Table A3. 2. Comparison of small intestine and *in vitro* cell lines; 'y' indicates transporter and enzyme expression; Bold text indicates high expression; italic indicates moderate, normal text indicates low expression according to the literature from chapter 9

Cell line	**Small intestine**	**Caco-2**	**MDCK**	**References**
Species	Human	Human	Canine	(Volpe, 2008, Cho *et al*., 1989)
Tissue	Small intestine	Colon adenocarcinoma	Kidney	(Volpe, 2008, Cho *et al*., 1989)
Culture time (days)	N/A	21[a]	3-5	(Volpe, 2008, Cho *et al*., 1989)
TEER ($\Omega.cm^2$)	25-40	300-500	Parental MDCK <200 MDCK I >1000 MDCK II <300 MDCK-MDR1 1000-10000	(Dukes *et al*., 2011, Braun *et al*., 2000, Volpe, 2008, Putnam *et al*., 2002, Quan *et al*., 2012, Balimane and Chong, 2005, Irvine *et al*., 1999, Soldner *et al*., 2000, Cho *et al*., 1989)
Unstirred water layer (UWL) thickness (μm)	<100	1000-2500	NF	(Lennernas, 1998, Hilgers *et al*., 1990, Karlsson and Artursson, 1992, Hidalgo *et al*., 1991)
Transporter expression	Transporter expression is highest in small intestine (area dependent) compared to cell lines in most cases. Transport expression is lower in MDCK cells compared to Caco-2 cells(Braun *et al*., 2000)			
MDR1(P-gp)	*y*	*y*	y y MDCK II *y* MDCK-MDR1	(Braun *et al*., 2000, Kuteykin-Teplyakov *et al*., 2010, Balimane *et al*., 2007, Seithel *et al*., 2006, Maubon *et al*., 2007, Hayeshi *et al*., 2008, Quan *et al*., 2012, Di *et al*., 2011)
MDR3	y	y	y MDCK II	(Quan *et al*., 2012, Hilgendorf *et al*., 2007, Hayeshi *et al*., 2008)
HPT1	**y**	**y**	NF	(Hilgendorf *et al*., 2007, Behrens *et al*., 2004, Sun *et al*., 2002, Hayeshi *et al*., 2008)
PEPT1	**y**	*y*	ND MDCK II	(Hilgendorf *et al*., 2007, Englund *et al*., 2006, Balimane *et al*., 2007, Seithel *et al*., 2006, Sun *et al*., 2002, Maubon *et al*., 2007, Hayeshi *et al*., 2008, Quan *et al*., 2012)
PEPT2	ND	ND	y MDCK II	(Leibach and Ganapathy, 1996, Balimane *et al*., 2007, Hilgendorf *et al*., 2007, Hayeshi *et al*., 2008, Quan *et al*., 2012)
OCTN1	y	y	NF	(Sun *et al*., 2002, Maubon *et al*., 2007, Hilgendorf *et al*., 2007, Hayeshi *et al*., 2008)
OCTN2	*y*	y	NF	(Hilgendorf *et al*., 2007, Volpe, 2008, Seithel *et al*., 2006, Maubon *et al*., 2007, Hayeshi *et al*., 2008)
MCT1	*y*	*y*	y MDCK II	(Deora *et al*., 2005, Hilgendorf *et al*., 2007, Seithel *et al*., 2006, Maubon *et al*., 2007)
MCT2	ND	y	NF	(Morris and Felmlee, 2008, Halestrap and Meredith, 2004, Lin *et al*., 1998)
MCT3	y	y	ND MDCK II	(Sun *et al*., 2002, Deora *et al*., 2005, Hayeshi *et al*., 2008)
MCT4	y	*y*	y MDCK II	(Deora *et al*., 2005, Morris and Felmlee, 2008, Halestrap and Meredith, 2004, Gill *et al*., 2005, Hayeshi *et al*., 2008)
MCT5	*y*	y	NF	(Hilgendorf *et al*., 2007, Hayeshi *et al*., 2008)
IBAT	*y*	*y*	NF	(Hilgendorf *et al*., 2007, Putnam *et al*., 2002, Hayeshi *et al*., 2008)
OAT1	ND	ND	ND MDCK II	(Quan *et al*., 2012, Zalups and Ahmad, 2005, Hilgendorf *et al*., 2007, Seithel *et al*., 2006, Hayeshi *et al*., 2008)
OAT2	y	y	NF	(Hilgendorf *et al*., 2007, Seithel *et al*., 2006)
OAT3	ND	ND	NF	(Hilgendorf *et al*., 2007, Hayeshi *et al*., 2008)
OAT4	ND	y	NF	(Hilgendorf *et al*., 2007, Whitley *et al*., 2006, Hayeshi *et al*., 2008)
OATP-A	y	y	ND MDCK II	(Hilgendorf *et al*., 2007, Maubon *et al*., 2007, Glaeser *et al*., 2007, Quan *et al*., 2012, Goh *et al*., 2002)

Cell line	Small intestine	Caco-2	MDCK	References
OATP-B	y	**y**	NF	(Englund *et al*., 2006, Seithel *et al*., 2006, Maubon *et al*., 2007, Glaeser *et al*., 2007, Hayeshi *et al*., 2008)
OATP-C	ND	ND	ND MDCK II ND*	(Hilgendorf *et al*., 2007, Goh *et al*., 2002, Hayeshi *et al*., 2008)
OATP-D	y	y	NF	(Hilgendorf *et al*., 2007, Hayeshi *et al*., 2008, Sai *et al*., 2006)
OATP-E	y	y	NF	(Hilgendorf *et al*., 2007, Hayeshi *et al*., 2008, Sai *et al*., 2006)
OATP-F	y	ND	NF	(Hilgendorf *et al*., 2007, Hayeshi *et al*., 2008)
OATP-H	y	y	ND* ND MDCK II	(Hilgendorf *et al*., 2007, Hayeshi *et al*., 2008, Mikkaichi *et al*., 2004, Kuo *et al*., 2012)
OATP8	y	ND	ND MDCK II	(Hilgendorf *et al*., 2007, Glaeser *et al*., 2007, Hayeshi *et al*., 2008, Quan *et al*., 2012)
BCRP	**y**	*y*	ND* ND MDCK II	(Quan *et al*., 2012, Englund *et al*., 2006, Xia *et al*., 2005, Seithel *et al*., 2006, Maubon *et al*., 2007, Hayeshi *et al*., 2008, Kuteykin-Teplyakov *et al*., 2010, Di *et al*., 2011)
MRP1	y	y	y* y MDCK II y MDCK-MDR1	(Maubon *et al*., 2007, Hayeshi *et al*., 2008, Hilgendorf *et al*., 2007, Di *et al*., 2011, Goh *et al*., 2002, Kuteykin-Teplyakov *et al*., 2010)
MRP2	**y**	**y**	ND* y MDCK II y MDCK-MDR1	(Englund *et al*., 2006, Quan *et al*., 2012, Seithel *et al*., 2006, Maubon *et al*., 2007, Hayeshi *et al*., 2008, Hilgendorf *et al*., 2007, Goh *et al*., 2002, Kuteykin-Teplyakov *et al*., 2010)
MRP3	*y*	*y*	y MDCK II	(Hilgendorf *et al*., 2007, Quan *et al*., 2012, Seithel *et al*., 2006, Maubon *et al*., 2007)
MRP5	y	y	y MDCK II y MDCK-MDR1	(Quan *et al*., 2012, Sun *et al*., 2002, Hilgendorf *et al*., 2007, Di *et al*., 2011, Kuteykin-Teplyakov *et al*., 2010)
MRP6	y	y	NF	(Maubon *et al*., 2007)
OCT1	y	y	NF	(Seithel *et al*., 2006, Hilgendorf *et al*., 2007, Maubon *et al*., 2007, Hayeshi *et al*., 2008)
OCT2	ND	ND	y MDCK II	(Shu *et al*., 2001, Quan *et al*., 2012, Hilgendorf *et al*., 2007, Hayeshi *et al*., 2008)
OCT3	y	y	ND MDCK II	(Seithel *et al*., 2006, Maubon *et al*., 2007, Hayeshi *et al*., 2008, Quan *et al*., 2012)
ENT1	y	y	y MDCK II	(Quan *et al*., 2012, Sun *et al*., 2002, Hayeshi *et al*., 2008, Hammond *et al*., 2004)
LNAA (LAT1/LAT2)	y	y	y*	(Putnam *et al*., 2002, Rossier *et al*., 1999)
PHT1	y	y	y MDCK II	(Quan *et al*., 2012, Herrera-Ruiz *et al*., 2001)
LRP	*y*	NF	NF	(Hilgendorf *et al*., 2007, Taipalensuu *et al*., 2001)
Cycloph A	**y**	NF	NF	(Hilgendorf *et al*., 2007)
CNT3	ND	*y*	ND*	(Hilgendorf *et al*., 2007, Hayeshi *et al*., 2008, Errasti-Murugarren *et al*., 2007)
Enzyme expression	Enzymes expressed in the small intestine have higher abundance compared to the cell lines.			
CYP3A4	**y**	y (very low)	ND*	(Braun *et al*., 2000, Sun *et al*., 2002, Paine *et al*., 2006, Borlak and Zwadlo, 2003, Kwatra *et al*., 2012)
CYP2D6	y	y (very low)	NF	(Paine *et al*., 2006, Hayeshi *et al*., 2008, Borlak and Zwadlo, 2003)
CYP2C9	**y**	*y*	NF	(Sun *et al*., 2002, Paine *et al*., 2006, Hayeshi *et al*., 2008, Borlak and Zwadlo, 2003)
CYP2C19	*y*	*y*	NF	(Sun *et al*., 2002, Paine *et al*., 2006, Hayeshi *et al*., 2008, Borlak and Zwadlo, 2003)
CYP1A1	y	y	NF	(Paine *et al*., 2006, Borlak and Zwadlo, 2003)
CYP2J2	*y*	y	ND MDCK II	(Sun *et al*., 2002, Paine *et al*., 2006, Borlak and Zwadlo, 2003, Quan *et al*., 2012)
CYP3A5	y	y	NF	(Paine *et al*., 2006, Hayeshi *et al*., 2008, Borlak and Zwadlo, 2003)

Cell line	Small intestine	Caco-2	MDCK	References
UGT	y	y	NF	(Sun *et al*., 2002)
GST	y	y	y*	(Sun *et al*., 2002, Volpe, 2008, Bohets *et al*., 1996)
ST	y	y	y MDCK II	(Sun *et al*., 2002, Volpe, 2008, Ng *et al*., 2003)
AT	y	y	NF	(Sun *et al*., 2002)
AcyT	y	y	NF	(Sun *et al*., 2002)

[a]when P-gp expression is at its maximum(Braun *et al*., 2000); NF- not found in the literature; ND - not detected in experimental assay; MDR - multi drug resistant protein; HPT1 - human oligopeptide transporter; PEPT- peptide transporter; OCTN - organic cation transporter; MCT- monocarboxylate transporters; IBAT - ileal sodium-dependent bile acid transporter/ intestinal bile acid transporter; OAT - organic anion transporter; OATP - Organic anion-transporting polypeptide; BCRP - Breast cancer resistance protein; MRP - multidrug resistance associated protein; OCT - organic cation transporter; ENT - equilibrative nucleoside transporter; LNAA/ LAT - Large neutral amino acids; LRP - lipoprotein transporter; Cycloph A - cyclophilin A/ Peptidyl-prolyl Isomerase A, CNT - Concentrative nucleoside transporter; CYP - Cytochrome P450 enzyme; UGT- Uridine 5'-diphospho-glucuronosyltransferase; GST - glutathione S-transferase; ST- sulfotransferase; AT - N-acetyltransferase; AcyT – acyltrtransferase

*Note: A generic term MDCK is used unless specific information regarding the transporter in a specific strain is known. Therefore MDCK could indicate that the parietal MDCK has been used or that the strain was not specified in the study

References for Table A3.2

BALIMANE, P. V. & CHONG, S. 2005. Cell culture-based models for intestinal permeability: a critique. *Drug Discovery Today,* 10**,** 335-343.

BALIMANE, P. V., CHONG, S., PATEL, K., QUAN, Y., TIMOSZYK, J., HAN, Y. H., WANG, B., VIG, B. & FARIA, T. N. 2007. Peptide transporter substrate identification during permeability screening in drug discovery: Comparison of transfected MDCK-hPepT1 cells to Caco-2 cells. *Archives of Pharmacal Research,* 30**,** 507-518.

BEHRENS, I., KAMM, W., DANTZIG, A. H. & KISSEL, T. 2004. Variation of peptide transporter (PepT1 expression in Caco-2 cells as a function and HPT1) of cell origin. *Journal of Pharmaceutical Sciences,* 93**,** 1743-1754.

BOHETS, H. H., NOUWEN, E. J., DEBROE, M. E. & DIERICKX, P. J. 1996. The cytosolic glutathione S-transferase isoenzymes in the dog kidney cortex as compared with the corresponding MDCK renal cell line. *Biochimica Et Biophysica Acta-Molecular Cell Research,* 1311**,** 93-101.

BORLAK, J. & ZWADLO, C. 2003. Expression of drug-metabolizing enzymes, nuclear transcription factors and ABC transporters in Caco-2 cells. *Xenobiotica,* 33**,** 927-943.

BRAUN, A., HAMMERLE, S., SUDA, K., ROTHEN-RUTISHAUSER, B., GUNTHERT, M., KRAMER, S. D. & WUNDERLI-ALLENSPACH, H. 2000. Cell cultures as tools in biopharmacy. *European Journal of Pharmaceutical Sciences,* 11**,** S51-S60.

CHO, M. J., THOMPSON, D. P., CRAMER, C. T., VIDMAR, T. J. & SCIESZKA, J. F. 1989. The madin darby canine kidney (MDCK) epithelial-cell monolayer as a model cellular-transport barrier. *Pharmaceutical Research,* 6**,** 71-77.

DEORA, A. A., PHILP, N., HU, J., BOK, D. & RODRIGUEZ-BOULAN, E. 2005. Mechanisms regulating tissue-specific polarity of monocarboxylate transporters and their chaperone CD147 in kidney and retinal epithelia. *Proceedings of the National Academy of Sciences of the United States of America,* 102**,** 16245-16250.

DI, L., WHITNEY-PICKETT, C., UMLAND, J. P., ZHANG, H., ZHANG, X., GEBHARD, D. F., LAI, Y. R., FEDERICO, J. J., DAVIDSON, R. E., SMITH, R., REYNER, E. L., LEE, C., FENG, B., ROTTER, C., VARMA, M. V., KEMPSHALL, S., FENNER, K., EL-KATTAN, A. F., LISTON, T. E. & TROUTMAN, M. D. 2011. Development of a New Permeability Assay Using Low-Efflux MDCKII Cells. *Journal of Pharmaceutical Sciences,* 100**,** 4974-4985.

DUKES, J. D., WHITLEY, P. & CHALMERS, A. D. 2011. The MDCK variety pack: choosing the right strain. *BMC Cell Biology,* 12.

ENGLUND, G., RORSMAN, F., RONNBLOM, A., KARLBOM, U., LAZOROVA, L., GRASJO, J., KINDMARK, A. & ARTURSSON, P. 2006. Regional levels of drug transporters along the human intestinal tract: Co-expression of ABC and SLC transporters and comparison with Caco-2 cells. *European Journal of Pharmaceutical Sciences,* 29**,** 269-277.

ERRASTI-MURUGARREN, E., PASTOR-ANGLADA, M. & CASADO, F. J. 2007. Role of CNT3 in the transepithelial flux of nucleosides and nucleoside-derived drugs. *Journal of Physiology,* 582**,** 1249-1260.

GILL, R. K., SAKSENA, S., ALREFAI, W. A., SARWAR, Z., GOLDSTEIN, J. L., CARROLL, R. E., RAMASWAMY, K. & DUDEJA, P. K. 2005. Expression and membrane localization of MCT isoforms along the length of the human intestine. *American Journal of Physiology-Cell Physiology,* 289**,** C846-C852.

GLAESER, H., BAILEY, D. G., DRESSER, G. K., GREGOR, J. C., SCHWARZ, U. I., MCGRATH, J. S., JOLICOEUR, E., LEE, W., LEAKE, B. F., TIRONA, R. G. & KIM, R. B. 2007. Intestinal drug transporter expression and the impact of grapefruit juice in humans. *Clinical Pharmacology & Therapeutics,* 81**,** 362-370.

GOH, L. B., SPEARS, K. J., YAO, D. G., AYRTON, A., MORGAN, P., WOLF, C. R. & FRIEDBERG, T. 2002. Endogenous drug transporters in in vitro and in vivo models for the prediction of drug disposition in man. *Biochemical Pharmacology,* 64**,** 1569-1578.

HALESTRAP, A. P. & MEREDITH, D. 2004. The SLC16 gene family - from monocarboxylate transporters (MCTs) to aromatic amino acid transporters and beyond. *Pflugers Archiv-European Journal of Physiology,* 447**,** 619-628.

HAMMOND, J. R., STOLK, M., ARCHER, R. G. E. & MCCONNELL, K. 2004. Pharmacological analysis and molecular cloning of the canine equilibrative nucleoside transporter 1. *European Journal of Pharmacology,* 491**,** 9-19.

HAYESHI, R., HILGENDORF, C., ARTURSSON, P., AUGUSTIJNS, P., BRODIN, B., DEHERTOGH, P., FISHER, K., FOSSATI, L., HOVENKAMP, E., KORJAMO, T., MASUNGI, C., MAUBON, N., MOLS, R., MULLERTZ, A., MONKKONEN, J., O'DRISCOLL, C., OPPERS-TIEMISSEN, H. M., RAGNARSSON, E. G. E., ROOSEBOOM, M. & UNGELL, A. L. 2008. Comparison of drug transporter gene expression and functionality in Caco-2 cells from 10 different laboratories. *European Journal of Pharmaceutical Sciences,* 35**,** 383-396.

HERRERA-RUIZ, D., WANG, Q., COOK, T. J., KNIPP, G. T., GUDMUNDSSON, O. S., SMITH, R. L. & FARIA, T. N. 2001. Spatial expression patterns of peptide transporters in the human and rat gastrointestinal tracts, Caco-2 in vitro cell culture model, and multiple human tissues. *AAPS Pharmsci,* 3**,** art. no.-9.

HIDALGO, I. J., HILLGREN, K. M., GRASS, G. M. & BORCHARDT, R. T. 1991. Characterization of the unstirred water layer in caco-2 cell monolayers using a novel diffusion apparatus. *Pharmaceutical Research,* 8**,** 222-227.

HILGENDORF, C., AHLIN, G., SEITHEL, A., ARTURSSON, P., UNGELL, A. L. & KARLSSON, J. 2007. Expression of thirty-six drug transporter genes in human intestine, liver, kidney, and organotypic cell lines. *Drug Metabolism and Disposition,* 35**,** 1333-1340.

HILGERS, A. R., CONRADI, R. A. & BURTON, P. S. 1990. Caco-2 cell monolayers as a model for drug transport across the intestinal-mucosa. *Pharmaceutical Research,* 7**,** 902-910.

IRVINE, J. D., TAKAHASHI, L., LOCKHART, K., CHEONG, J., TOLAN, J. W., SELICK, H. E. & GROVE, J. R. 1999. MDCK (Madin-Darby canine kidney) cells: A tool for membrane permeability screening. *Journal of Pharmaceutical Sciences,* 88**,** 28-33.

KARLSSON, J. & ARTURSSON, P. 1992. A new diffusion chamber system for the determination of drug permeability coefficients across the human intestinal epithelium that are independent of the unstirred water layer. *Biochimica Et Biophysica Acta,* 1111**,** 204-210.

KUO, K.-L., ZHU, H., MCNAMARA, P. J. & LEGGAS, M. 2012. Localization and Functional Characterization of the Rat Oatp4c1 Transporter in an In Vitro Cell System and Rat Tissues. *Plos One,* 7.

KUTEYKIN-TEPLYAKOV, K., LUNA-TORTOS, C., AMBROZIAK, K. & LOSCHER, W. 2010. Differences in the expression of endogenous efflux transporters in MDR1-transfected versus wildtype cell lines affect P-glycoprotein mediated drug transport. *British Journal of Pharmacology,* 160**,** 1453-1463.

KWATRA, D., BUDDA, B., VADLAPUDI, A. D., VADLAPATLA, R. K., PAL, D. & MITRA, A. K. 2012. Transfected MDCK Cell Line with Enhanced Expression of CYP3A4 and P-Glycoprotein as a Model To Study Their Role in Drug Transport and Metabolism. *Molecular Pharmaceutics,* 9**,** 1877-1886.

LEIBACH, F. H. & GANAPATHY, V. 1996. Peptide transporters in the intestine and the kidney. *Annual Review of Nutrition,* 16**,** 99-119.

LENNERNAS, H. 1998. Human intestinal permeability. *Journal of Pharmaceutical Sciences,* 87**,** 403-410.

LIN, R. Y., VERA, J. C., CHAGANTI, R. S. K. & GOLDE, D. W. 1998. Human monocarboxylate transporter 2 (MCT2) is a high affinity pyruvate transporter. *Journal of Biological Chemistry,* 273**,** 28959-28965.

MAUBON, N., LE VEE, M., FOSSATI, L., AUDRY, M., LE FERREC, E., BOLZE, S. & FARDEL, O. 2007. Analysis of drug transporter expression in human intestinal Caco-2 cells by real-time PCR. *Fundamental & Clinical Pharmacology,* 21**,** 659-663.

MIKKAICHI, T., SUZUKI, T., ONOGAWA, T., TANEMOTO, M., MIZUTAMARI, H., OKADA, M., CHAKI, T., MASUDA, S., TOKUI, T., ETO, N., ABE, M., SATOH, F., UNNO, M., HISHINUMA, T., INUI, K., ITO, S., GOTO, J. & ABE, T. 2004. Isolation and characterization of a digoxin transporter and its rat homologue expressed in the kidney. *Proceedings of the National Academy of Sciences of the United States of America,* 101**,** 3569-3574.

MORRIS, M. E. & FELMLEE, M. A. 2008. Overview of the proton-coupled MCT (SLC16A) family of transporters: Characterization, function and role in the transport of the drug of abuse gamma-hydroxybutyric acid. *AAPS Journal,* 10**,** 311-321.

NG, K. H., LIM, B. G. & WONG, K. P. 2003. Sulfate conjugating and transport functions of MDCK distal tubular cells. *Kidney International,* 63**,** 976-986.

PAINE, M. F., HART, H. L., LUDINGTON, S. S., HAINING, R. L., RETTIE, A. E. & ZELDIN, D. C. 2006. The human intestinal cytochrome P450 "pie". *Drug Metabolism and Disposition,* 34**,** 880-886.

PUTNAM, W. S., RAMANATHAN, S., PAN, L., TAKAHASHI, L. H. & BENET, L. Z. 2002. Functional characterization of monocarboxylic acid, large neutral amino acid, bile acid and peptide transporters, and P-glycoprotein in MDCK and Caco-2 cells. *Journal of Pharmaceutical Sciences,* 91**,** 2622-2635.

QUAN, Y., JIN, Y., FARIA, T. N., C. A. TILFORD, C. A., HE, A., WALL, D. A., SMITH, R. L. & VIG, B. S. 2012. Expression Profile of Drug and Nutrient Absorption Related Genes in Madin-Darby Canine Kidney (MDCK) Cells Grown under Differentiation Conditions. *Pharmaceutics,* 4**,** 314-333.

ROSSIER, G., MEIER, C., BAUCH, C., SUMMA, V., SORDAT, B., VERREY, F. & KUHN, L. C. 1999. LAT2, a new basolateral 4F2hc/CD98-associated amino acid transporter of kidney and intestine. *Journal of Biological Chemistry,* 274**,** 34948-34954.

SAI, Y., KANEKO, Y., ITO, S., MITSUOKA, K., KATO, Y., TAMAI, I., ARTURSSON, P. & TSUJI, A. 2006. Predominant contribution of organic anion transporting polypeptide OATP-B (OATP2B1) to apical uptake of estrone-3-sulfate by human intestinal Caco-2 cells. *Drug Metabolism and Disposition,* 34**,** 1423-1431.

SEITHEL, A., KARLSSON, J., HILGENDORF, C., BJORQUIST, A. & UNGELL, A. L. 2006. Variability in mRNA expression of ABC- and SLC-transporters in human intestinal cells: Comparison between human segments and Caco-2 cells. *European Journal of Pharmaceutical Sciences,* 28**,** 291-299.

SHU, Y., BELLO, C. L., MANGRAVITE, L. M., FENG, B. & GIACOMINI, K. M. 2001. Functional characteristics and steroid hormone-mediated regulation of an organic cation transporter in Madin-Darby canine kidney cells. *Journal of Pharmacology and Experimental Therapeutics,* 299**,** 392-398.

SOLDNER, A., BENET, L. Z., MUTSCHLER, E. & CHRISTIANS, U. 2000. Active transport of the angiotensin-II antagonist losartan and its main metabolite EXP 3174 across MDCK-MDR1 and Caco-2 cell monolayers. *British Journal of Pharmacology,* 129**,** 1235-1243.

SUN, D. X., LENNERNAS, H., WELAGE, L. S., BARNETT, J. L., LANDOWSKI, C. P., FOSTER, D., FLEISHER, D., LEE, K. D. & AMIDON, G. L. 2002. Comparison of human duodenum and Caco-2 gene expression profiles for 12,000 gene sequences tags and correlation with permeability of 26 drugs. *Pharmaceutical Research,* 19**,** 1400-1416.

TAIPALENSUU, J., TORNBLOM, H., LINDBERG, G., EINARSSON, C., SJOQVIST, F., MELHUS, H., GARBERG, P., SJOSTROM, B., LUNDGREN, B. & ARTURSSON, P. 2001. Correlation of gene expression of ten drug efflux proteins of the ATP-binding cassette transporter family in normal human jejunum and in human intestinal epithelial Caco-2 cell monolayers. *Journal of Pharmacology and Experimental Therapeutics,* 299**,** 164-170.

VOLPE, D. A. 2008. Variability in Caco-2 and MDCK cell-based intestinal permeability assays. *Journal of Pharmaceutical Sciences,* 97**,** 712-725.

WHITLEY, A. C., SWEET, D. H. & WALLE, T. 2006. Site-specific accumulation of the cancer preventive dietary polyphenol ellagic acid in epithelial cells of the aerodigestive tract. *Journal of Pharmacy and Pharmacology,* 58**,** 1201-1209.

XIA, C. Q., LIU, N., YANG, D., MIWA, G. & GAN, L. S. 2005. Expression, localization, and functional characteristics of breast cancer resistance protein in Caco-2 cells. *Drug Metabolism and Disposition,* 33**,** 637-643.

ZALUPS, R. K. & AHMAD, S. 2005. Handling of cysteine S-conjugates of methylmercury in MDCK cells expressing human OAT1. *Kidney International,* 68**,** 1684-1699.

Comparison of absorption through MDCK and Caco2 cell lines using different mechanisms/ routes

The statistical significance of the regression lines (slope and intercept) of the absorption mechanisms highlighted in Table 9.3 (in chapter 9) was tested and the results are shown in Table A3.3 As some of the compounds undergo more than one absorption mechanism (Groups D, F, G and H in Table 9.3), the significance was also tested in either mechanism. For example, for group D (efflux and paracellular) this was tested with all these compounds classed as efflux or classed as paracellular. P values < 0.05 indicate that there is a significant difference between the intercept and slope of the lines of the absorption groups being studied.

Table A3. 3. Linear regression results and significance of the different absorption transport routes from Table 9.3 in Chapter 9 (First absorption mechanism is dominant for that set)

Model	Description	N	Slope	Slope p value	Intercept	Intercept p value
1*	Paracellular (B, D ,F ,H) Vs Transcellular (A)	11 83	-0.191 -0.874	0.0023**	-7.370 -0.688	n/a
2	Carrier mediated (C-H) Vs Passive (A, B)	96 89	0.777 0.880	0.297	-1.239 -0.663	0.525
3	Efflux (C, D, G, H) Vs Passive (A, B)	79 89	0.787 0.880	0.355	-1.194 -0.663	0.479
4	Influx (E, F, G, H) Vs Passive (A, B)	32 89	0.673 0.880	0.132	-1.899 -0.663	0.326
5	Efflux (C, D, G, H) Vs Influx (E, F)	79 17	0.787 0.750	0.867	-1.194 -1.376	0.895
6	Influx (E, F, G, H) Vs Efflux (C, D)	32 64	0.673 0.795	0.498	-1.899 -1.113	0.338
7	Influx (E, F) Vs Efflux (C, D)	17 64	0.750 0.795	0.844	-1.376 -1.113	0.915

* Paired t tests were carried out on the smaller group of permeability values to test the significance between the two cell lines of the group of compounds. In addition paired t tests were also carried out for all the groups in models 1-7 to test for significance between the cell lines. All tests revealed not significant differences between the two cell lines.

**indicates statistical significance between the groups

Figures A3.1-7: Figures showing the relationship between the different transport mechanisms (defined in Table 9.3 and Table A3.3) between the 185 compounds for caco-2 and MDCK permeability

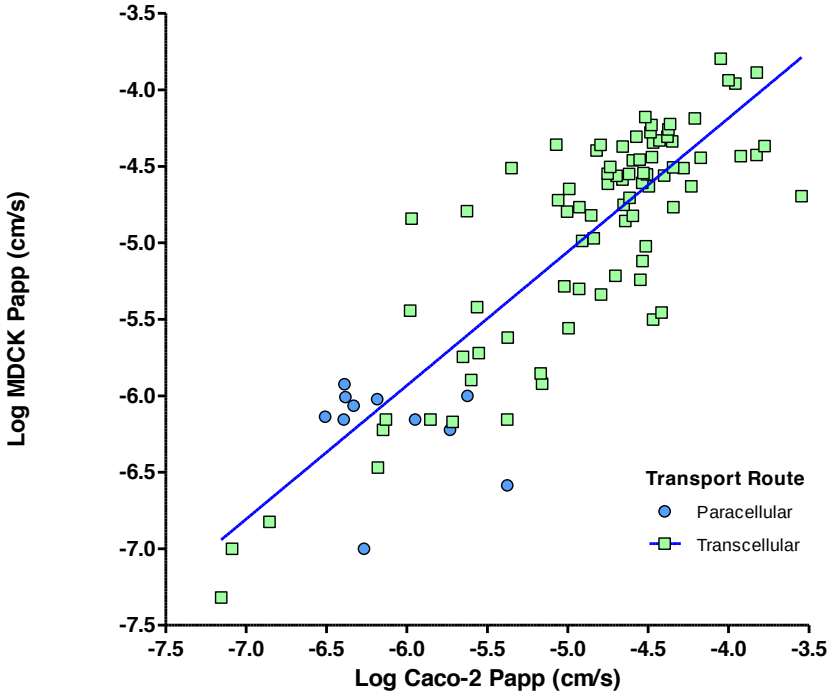

Figure A3. 1. Comparison of regression between compounds transcellular and paracellular (model 1 Table A3.3).

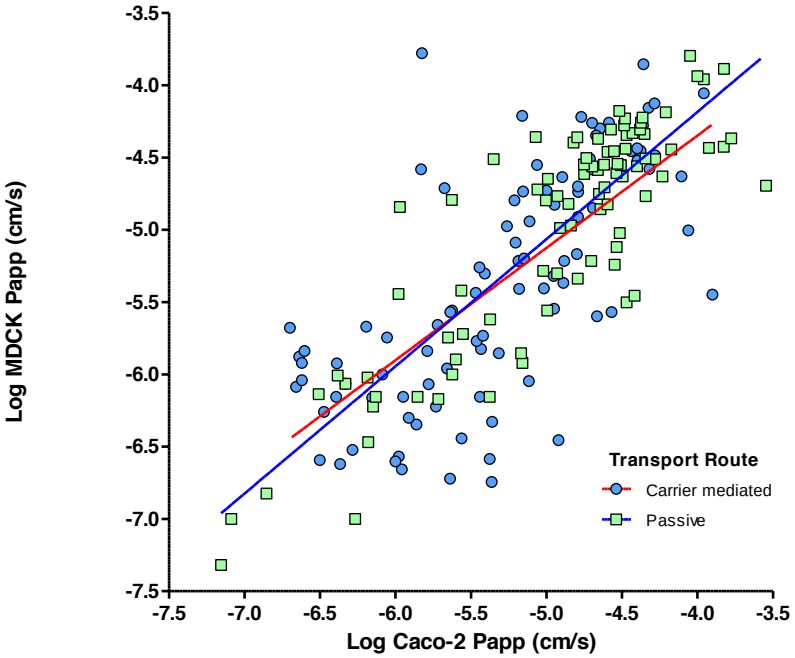

Figure A3. 2. Comparison of regression between carrier mediated transport and passive absorption compounds (model 2 Table A3.3)

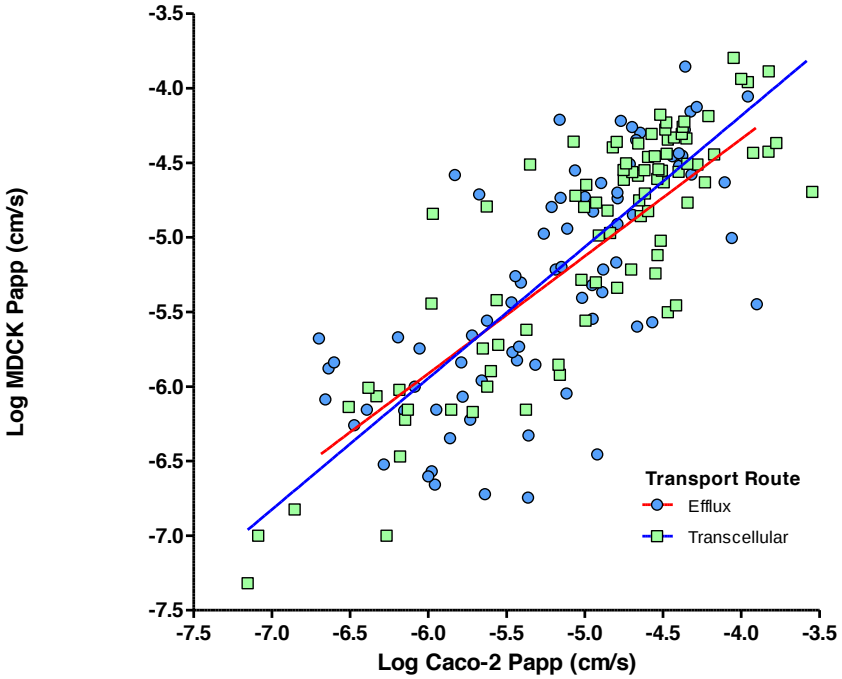

Figure A3. 3. Comparison of regression between compounds identified as carrier mediated efflux transport and passive absorption (model 3 Table A3.3)

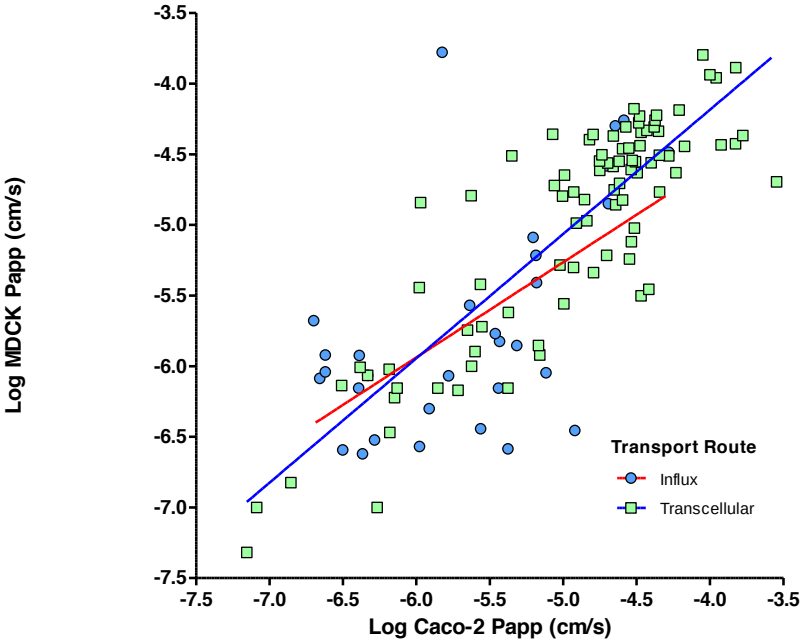

Figure A3. 4. Comparison of regression between compounds identified as carrier mediated influx transport and passive absorption (model 4 Table A3.3)

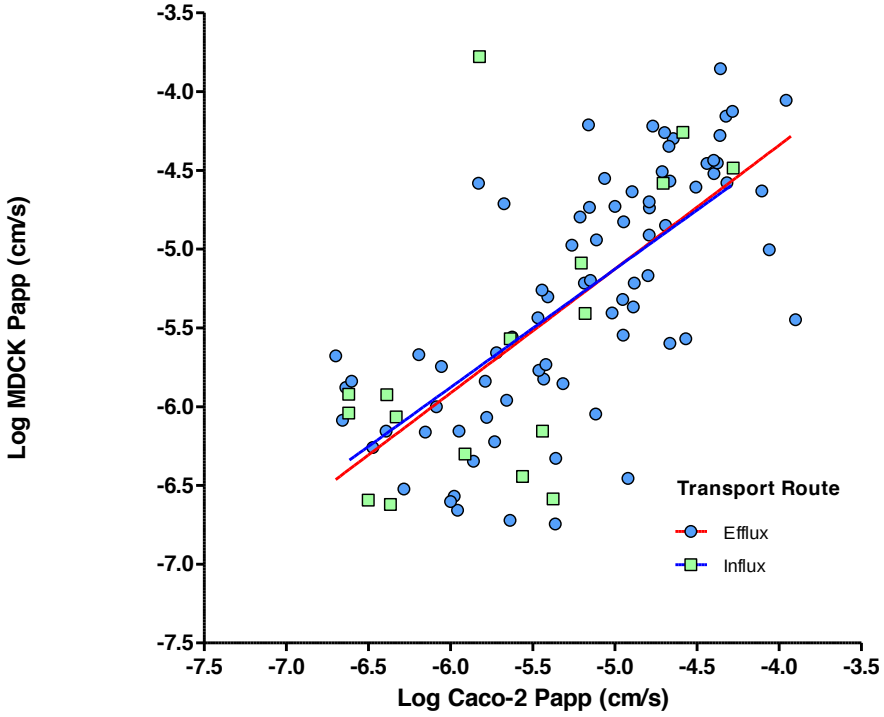

Figure A3. 5. Comparison of regression between carrier mediated efflux and influx transport (model 5 Table A3.3)

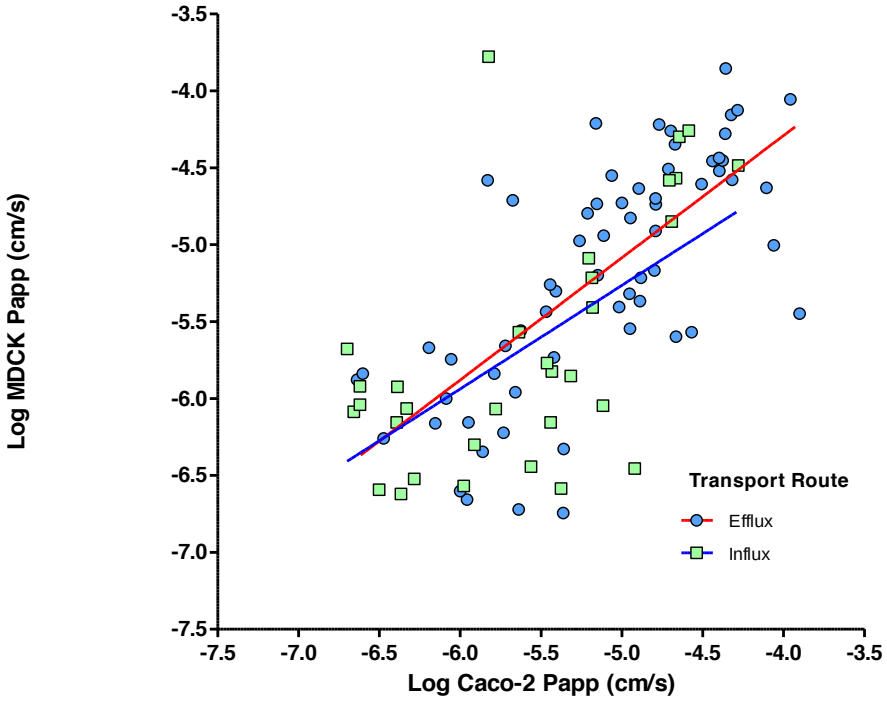

Figure A3. 6. Comparison of regression between carrier mediated efflux and influx transport (groups C-F Table 9.3) Model 6 in Table A3.3

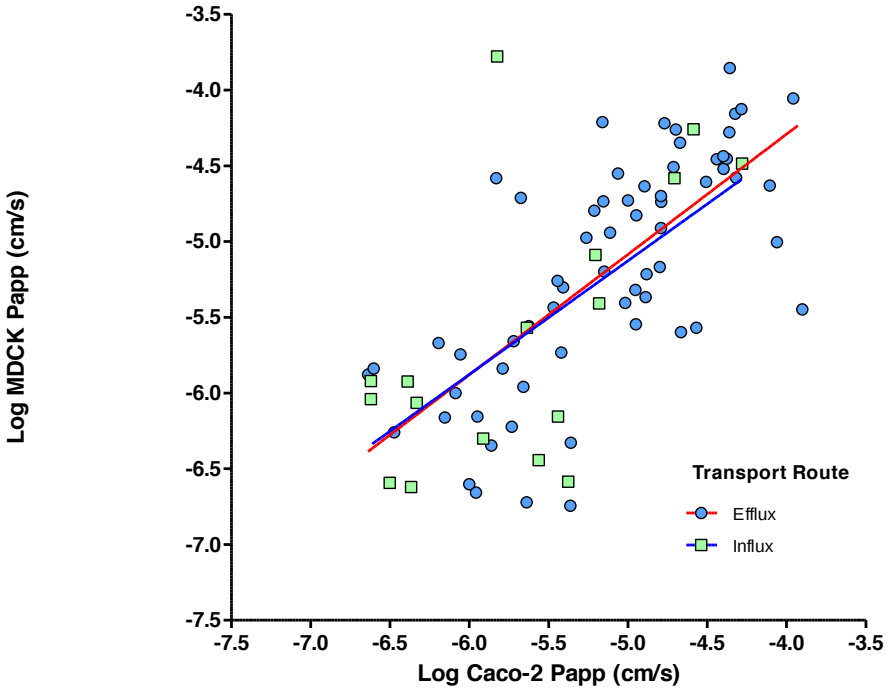

Figure A3. 7. Comparison of regression between carrier mediated influx (groups E, F Table 9.3) and efflux transport (groups C,D Table 9.3) Model 7 in Table A3.3

Table A3. 4. Potential Outliers in Sections of A and B in Figure 9.2 in chapter 9

Outlier type	Name	Therapeutic indication/Class	Carrier mediated	Gut met	Water solubility	Comments	References
POTENTIAL FALSE NEGATIVES High fraction absorbed Low cell permeability	Folinic Acid	Oncology	Influx, RFC		Sparingly soluble	Dose dependent absorption	(Matherly and Goldman, 2003, Balamurugan and Said, 2006, Dollery, 1999, O'Neil *et al*., 2001)
	Ribavirin	Antiviral	Influx, CNT2		Soluble		(Shugarts and Benet, 2009, O'Neil *et al*., 2001)
	Amoxicillin	Antibiotic	Influx, PEPT1		High solubility		(Estudante *et al*., 2013, O'Neil *et al*., 2001, Osol and Hoover, 1976) AQUASOL database 6th edition
	Loracarbef	Antibiotic	Influx, PEPT1		High solubility	Review article assigned compound into highly soluble group	(Hu *et al*., 1994, Benet *et al*., 2011)
	Baclofen	Gamma-aminobutyric acid agonist	Influx, Amino acid transporter		Slightly soluble		(Thwaites *et al*., 2000) Drugs@FDA, Product label
	Lamivudine	Antiviral	Influx, OCT3, Efflux, BCRP		High solubility		(Estudante *et al*., 2013, Minuesa *et al*., 2009, O'Neil *et al*., 2001)
	Floxuridine	Oncology	CNT3		High solubility	Possible influx substrate CNT3 nucleosides specificity however abundance of apical side of membrane is not certain	(Ritzel *et al*., 2001, Osol and Hoover, 1976) US EPA; Estimation Program Interface (EPI) Suite. Version 3.12
	Acetazolamide	Carbonic anhydrase inhibitor			Slightly soluble		(Budavari *et al*., 1996)
	Zolmitriptan	Serotonin receptor agonist	Influx, SERT Efflux, P-gp		High solubility	Possible influx substrate SERT serotonin specificity. Review article assigned compound into highly soluble group	(Benet *et al*., 2011, Martel *et al*., 2003)
	Diphenoxylate	Opioid			high/low solubility?	High logP ~ 5.66, interactions with the unstirred water layer in the *in vitro* assay? Conflicting sources state different solubilities	(Nguyen *et al*., 2012, WHO, 2006)

Table A3.4 (Cont): Potential Outliers in Sections of A and B in Figure 9.2 in chapter 9

Outlier type	Name	Therapeutic indication/Class	Carrier mediated	Gut met	Water solubility	Comments	References
POTENTIAL FALSE POSITVES Low fraction absorbed High cell permeability	Mebendazole	Anti parasitic	Efflux, P-gp	CBRs,	Low solubility		(Varma *et al*., 2010, Varma *et al*., 2005, Nishimuta *et al*., 2013)
	Albendazole	Anti parasitic		CYP3A4 and FMOs	Low solubility		(Redondo *et al*., 1999) Drugs@FDA Product information
	Cephalothin	Antibiotic		Esterases and endopeptidases	Moderate/High solubility	Classed as moderately soluble: >0.01mg/ml <1mg/ml Gut metabolism tested using Porcine	(Gozalbes and Pineda-Lucena, 2010, Sarti *et al*., 2011)
	Tacrolimus	Immunosuppressant	Efflux, P-gp	CYP3A4	Low solubility	Review article assigned compound into poorly soluble group	(Estudante *et al*., 2013, Benet *et al*., 2011, Varma *et al*., 2010, Hebert, 1997)
	Lovastatin	Statin		Yes, CYP3A4	Low solubility		(Varma *et al*., 2010, Jacobsen *et al*., 1999, Kato, 2008, O'Neil *et al*., 2001)
	Bromocriptine	Dopamine agonist	Efflux, P-gp	Yes, CYP3A4	Low solubility		(Vautier *et al*., 2006, Fagerholm, 2007, Yap and Chen, 2005) Norvartis product label

RFC: Reduced folate carrier; CNT2/3: Concentrative nucleoside transporter 2/3; PEPT1: Peptide transport protein; OCT3: Organic cation transporter; BCRP: Breast cancer resistance protein; SERT: Neuronal serotonin transporter; P-gp: P-Glycoprotein; CBR: Carbonyl reductases; FMO: Flavin-containing monooxygenases; CYP3A4: Cytochrome P450 3A4

References for Table A3.4

BALAMURUGAN, K. & SAID, H. M. 2006. Role of reduced folate carrier in intestinal folate uptake. *American Journal of Physiology-Cell Physiology,* 291**,** 22.

BENET, L. Z., BROCCATELLI, F. & OPREA, T. I. 2011. BDDCS applied to over 900 drugs. *The AAPS Journal,* 13**,** 519-47.

BUDAVARI, S., SMITH, A., KINNEARY, J., O'NEILL, M. & HECKELMAN, P. (eds.) 1996. *The Merck Index: An Encyclopedia of Chemicals, Drugs and Biologicals,* New Jersey: Merck & Co.

DOLLERY, C. 1999. *Therapeutic drugs*, Elsevier Science Health Science Division.

ESTUDANTE, M., MORAIS, J. G., SOVERAL, G. & BENET, L. Z. 2013. Intestinal drug transporters: an overview. *Advanced Drug Delivery Reviews,* 65**,** 1340-56.

FAGERHOLM, U. 2007. Prediction of human pharmacokinetics - gut-wall metabolism. *Journal of Pharmacy and Pharmacology,* 59**,** 1335-1343.

GOZALBES, R. & PINEDA-LUCENA, A. 2010. QSAR-based solubility model for drug-like compounds. *Bioorganic & Medicinal Chemistry,* 18**,** 7078-7084.

HEBERT, M. F. 1997. Contributions of hepatic and intestinal metabolism and P-glycoprotein to cyclosporine and tacrolimus oral drug delivery. *Advanced Drug Delivery Reviews,* 27**,** 201-214.

HU, M., CHEN, J., ZHU, Y., DANTZIG, A. H., STRATFORD, R. E., JR. & KUHFELD, M. T. 1994. Mechanism and kinetics of transcellular transport of a new beta-lactam antibiotic loracarbef across an intestinal epithelial membrane model system (Caco-2). *Pharmaceutical Research,* 11**,** 1405-13.

JACOBSEN, W., KIRCHNER, G., HALLENSLEBEN, K., MANCINELLI, L., DETERS, M., HACKBARTH, I., BANER, K., BENET, L. Z., SEWING, K. F. & CHRISTIANS, U. 1999. Small intestinal metabolism of the 3-hydroxy-3-methylglutary-coenzyme A reductase inhibitor lovastatin and comparison with pravastatin. *Journal of Pharmacology and Experimental Therapeutics,* 291**,** 131-139.

KATO, M. 2008. Intestinal first-pass metabolism of CYP3A4 substrates. *Drug Metabolism and Pharmacokinetics,* 23**,** 87-94.

MARTEL, F., MONTEIRO, R. & LEMOS, C. 2003. Uptake of serotonin at the apical and basolateral membranes of human intestinal epithelial (Caco-2) cells occurs through the neuronal serotonin transporter (SERT). *Journal of Pharmacology and Experimental Therapeutics,* 306**,** 355-62.

MATHERLY, L. H. & GOLDMAN, I. D. 2003. Membrane transport of folates. *Vitamins & Hormones,* 66**,** 403-456.

MINUESA, G., VOLK, C., MOLINA-ARCAS, M., GORBOULEV, V., ERKIZIA, I., ARNDT, P., CLOTET, B., PASTOR-ANGLADA, M., KOEPSELL, H. & MARTINEZ-PICADO, J. 2009. Transport of Lamivudine (-)-beta-L-2 ',3 '-Dideoxy-3 '-thiacytidine and High-Affinity Interaction of Nucleoside Reverse Transcriptase Inhibitors with Human Organic Cation Transporters 1, 2, and 3 (vol 329, pg 252, 2009). *Journal of Pharmacology and Experimental Therapeutics,* 329**,** 1187-1187.

NGUYEN, W., HOWARD, B. L., JENKINS, D. P., WULFF, H., THOMPSON, P. E. & MANALLACK, D. T. 2012. Structure-activity relationship exploration of Kv1.3 blockers based on diphenoxylate. *Bioorganic & Medicinal Chemistry Letters,* 22**,** 7106-9.

NISHIMUTA, H., NAKAGAWA, T., NOMURA, N. & YABUKI, M. 2013. Significance of reductive metabolism in human intestine and quantitative prediction of intestinal first-pass metabolism by cytosolic reductive enzymes. *Drug Metabolism and Disposition,* 41**,** 1104-11.

O'NEIL, M. J., SMITH, A., HECKELMAN, P. E. & BUDAVARI, S. (eds.) 2001. *The Merck Index: An Encyclopedia of Chemicals, Drugs and Biologicals,* New Jersey: Merck & Co.

OSOL, A. & HOOVER, J. E. (eds.) 1976. *Remington's Pharmaceutical Sciences,* Easton, Pennsylvania: Mack Publishing Co.

REDONDO, P. A., ALVAREZ, A. I., GARCIA, J. L., LARRODE, O. M., MERINO, G. & PRIETO, J. G. 1999. Presystemic metabolism of albendazole: experimental evidence of an efflux process of albendazole sulfoxide to intestinal lumen. *Drug Metabolism and Disposition,* 27**,** 736-40.

RITZEL, M. W., NG, A. M., YAO, S. Y., GRAHAM, K., LOEWEN, S. K., SMITH, K. M., RITZEL, R. G., MOWLES, D. A., CARPENTER, P., CHEN, X. Z., KARPINSKI, E., HYDE, R. J., BALDWIN, S. A., CASS, C. E. & YOUNG, J. D. 2001. Molecular identification and characterization of novel human and mouse concentrative Na+-nucleoside cotransporter proteins (hCNT3 and mCNT3) broadly selective for purine and pyrimidine nucleosides (system cib). *Journal of Biological Chemistry,* 276**,** 2914-27.

SARTI, F., BARTHELMES, J., IQBAL, J., HINTZEN, F. & BERNKOP-SCHNURCH, A. 2011. Intestinal enzymatic metabolism of drugs. *Journal of Pharmacology and Pharmacotherapeutics,* 63**,** 392-9.

SHUGARTS, S. & BENET, L. Z. 2009. The Role of Transporters in the Pharmacokinetics of Orally Administered Drugs. *Pharmaceutical Research,* 26**,** 2039-2054.

THWAITES, D. T., BASTERFIELD, L., MCCLEAVE, P. M. J., CARTER, S. M. & SIMMONS, N. L. 2000. Gamma-aminobutyric acid (GABA) transport across human intestinal epithelial (Caco-2) cell monolayers. *British Journal of Pharmacology,* 129**,** 457-464.

VARMA, M. V. S., OBACH, R. S., ROTTER, C., MILLER, H. R., CHANG, G., STEYN, S. J., EL-KATTAN, A. & TROUTMAN, M. D. 2010. Physicochemical Space for Optimum Oral Bioavailability: Contribution of Human Intestinal Absorption and First-Pass Elimination. *Journal of Medicinal Chemistry,* 53**,** 1098-1108.

VARMA, M. V. S., SATEESH, K. & PANCHAGNULA, R. 2005. Functional role of P-glycoprotein in limiting intestinal absorption of drugs: Contribution of passive permeability to P-glycoprotein mediated efflux transport. *Molecular Pharmaceutics,* 2**,** 12-21.

VAUTIER, S., LACOMBLEZ, L., CHACUN, H., PICARD, V., GIMENEZ, F., FARINOTTI, R. & FERNANDEZ, C. 2006. Interactions between the dopamine agonist, bromocriptine and the efflux protein, P-glycoprotein at the blood-brain barrier in the mouse. *European Journal of Pharmaceutical Sciences,* 27**,** 167-74.

WHO 2006. *The International Pharmacopoeia,* Geneva, World Health Organization.

YAP, C. W. & CHEN, Y. Z. 2005. Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines. *Journal of Chemical Information and Modeling,* 45**,** 982-92.

Figures A3.8-19. C&RT model analysis from Table 9.5 from Chapter 9

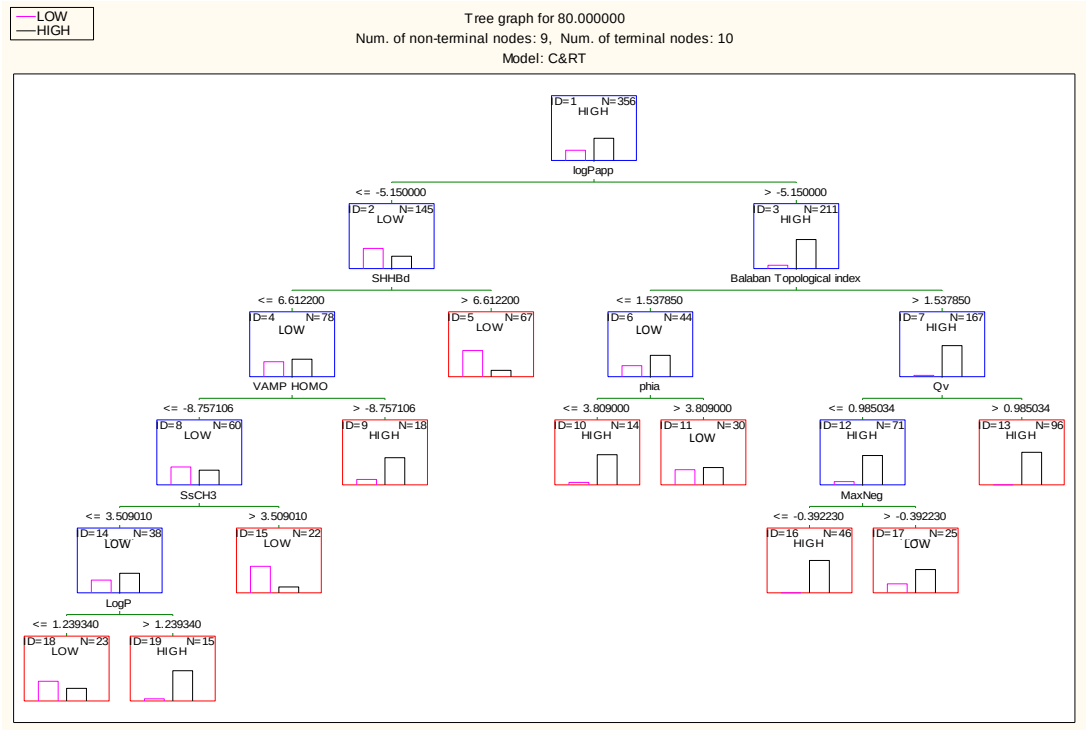

Figure A3. 8 Model 1; C&RT permeability model when higher misclassification costs of three and six were applied to reduce false positives for the highly and poorly permeability nodes.

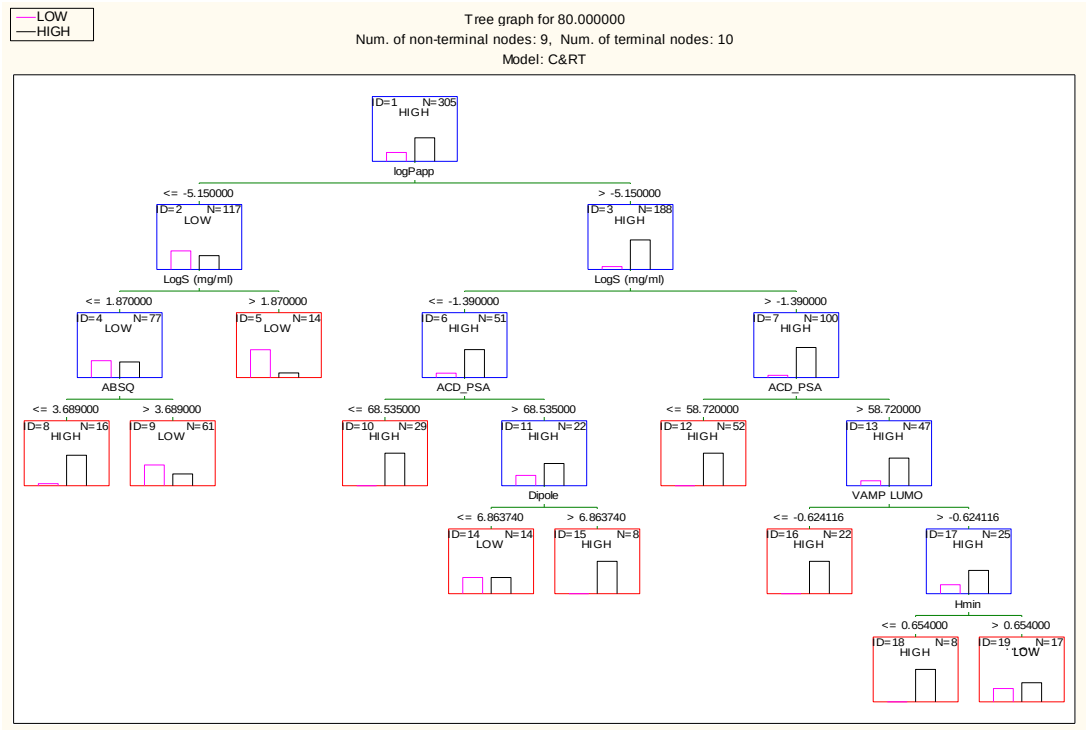

Figure A3. 9. Model 2; C&RT permeability and experimental solubility (logS mg/mL) model when higher misclassification costs of two and ten to reduce false positives were applied to high and low permeability compound nodes

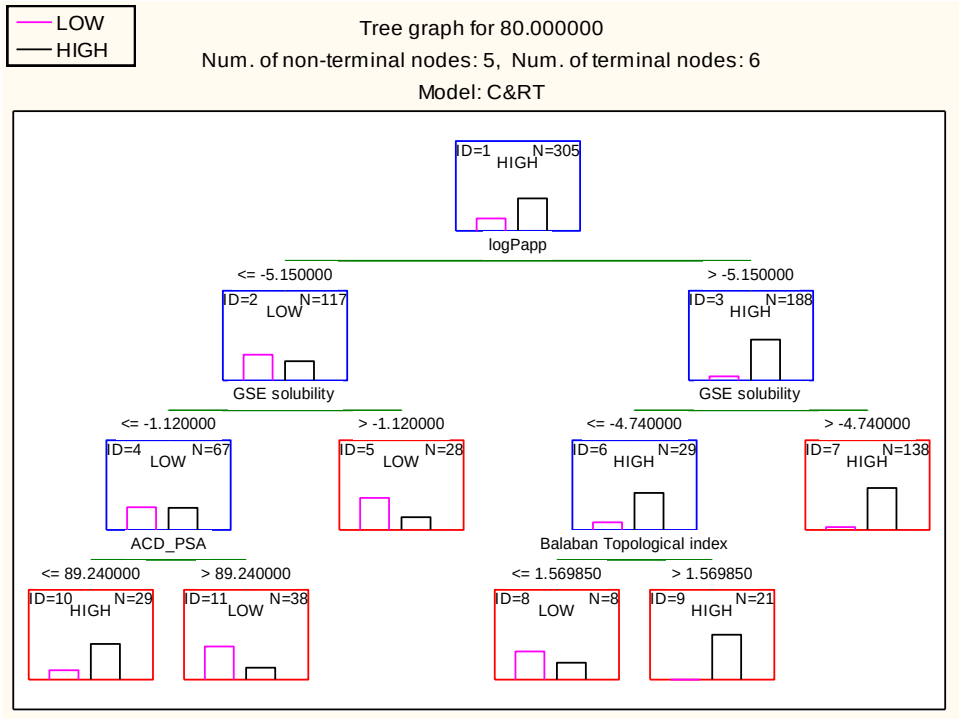

Figure A3. 10. Model 3; C&RT permeability and predicted solubility (GSE) model when higher misclassification costs of two to reduce false positives were applied to high GSE solubility node

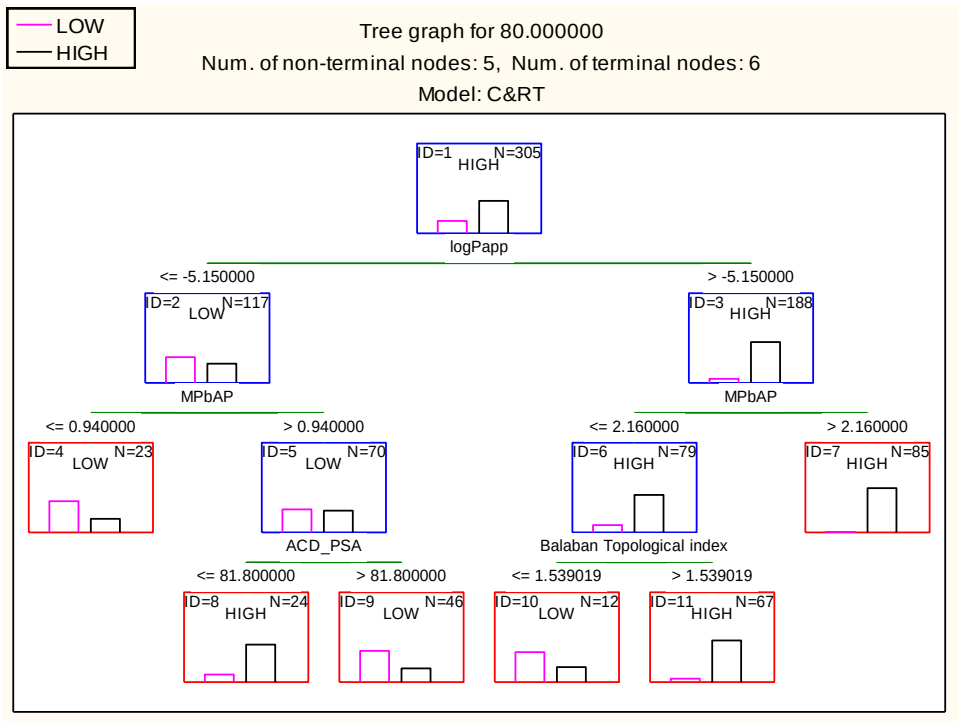

Figure A3. 11. Model 4; C&RT permeability and melting point based absorption potential (MPbAP) model when equal misclassification costs were applied to both nodes

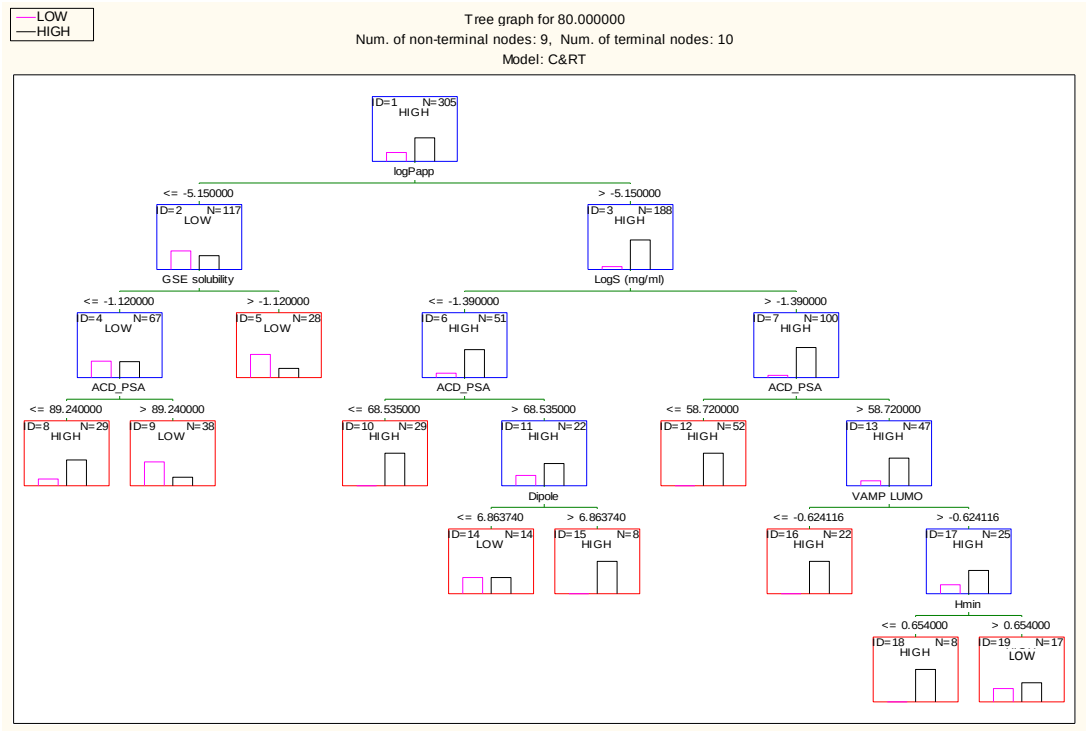

Figure A3. 12. Model 5; C&RT permeability, predicted solubility (GSE) and experimental solubility (LogS mg/mL) model when higher misclassification costs of two and ten to reduce false positives were applied to the high logS (mg/mL) node and low GSE solubility node respectively

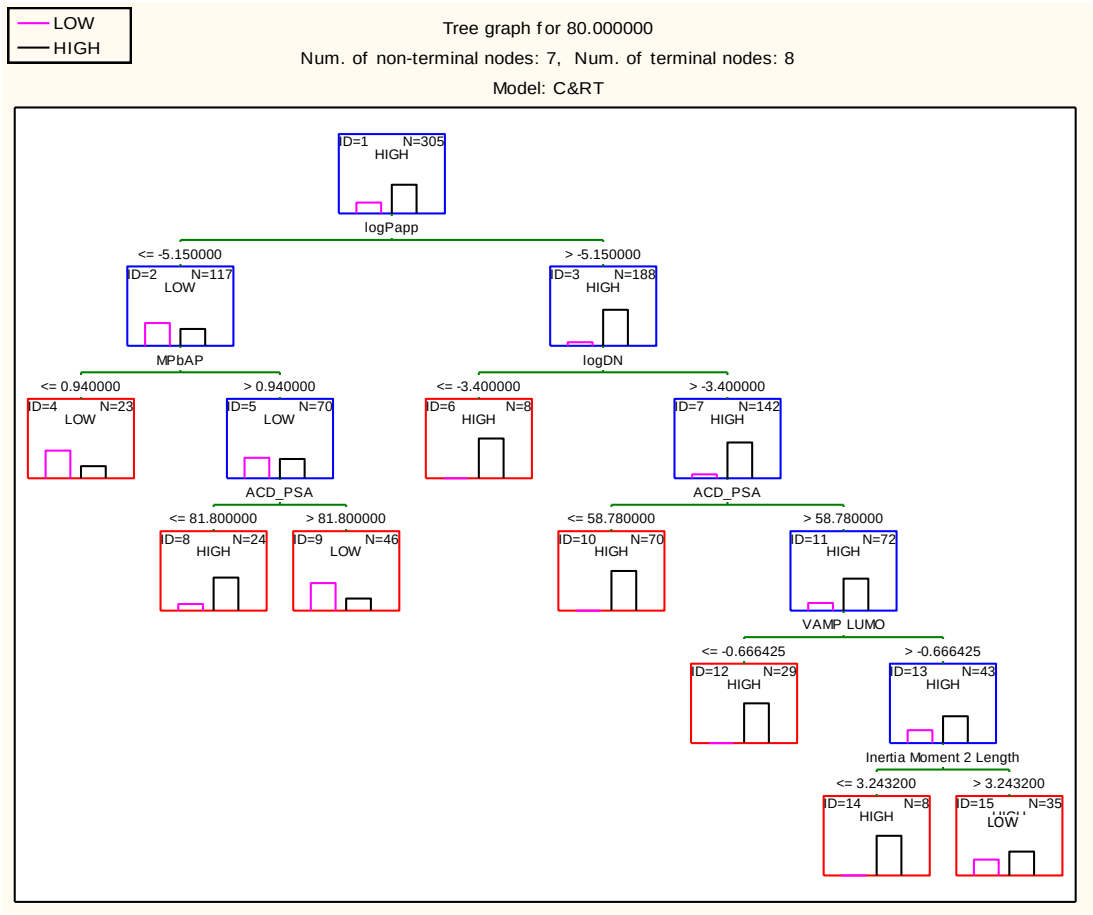

Figure A3. 13. Model 6; C&RT permeability, melting point based absorption potential (MPbAP) and log Dose number (LogDN) model when higher misclassification costs of two and ten to reduce false positives were applied to the high logDN node and low MPbAP node respectively

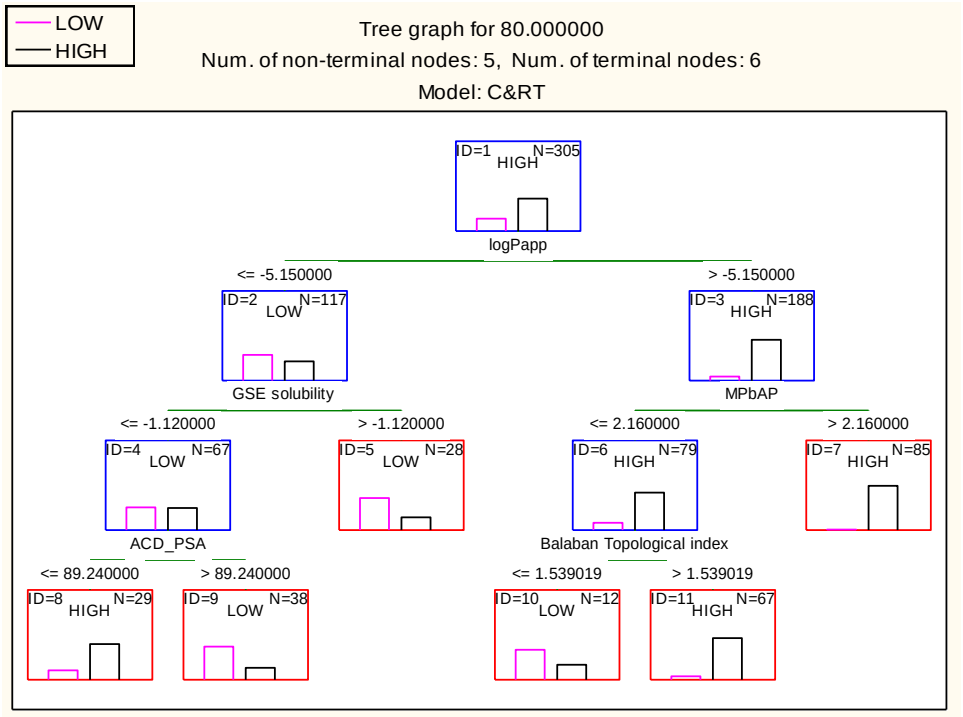

Figure A3. 14. Model 7; C&RT permeability predicted solubility (GSE) and melting point based absorption potential (MPbAP) model when higher misclassification costs of two were applied to the high MPbAP node only

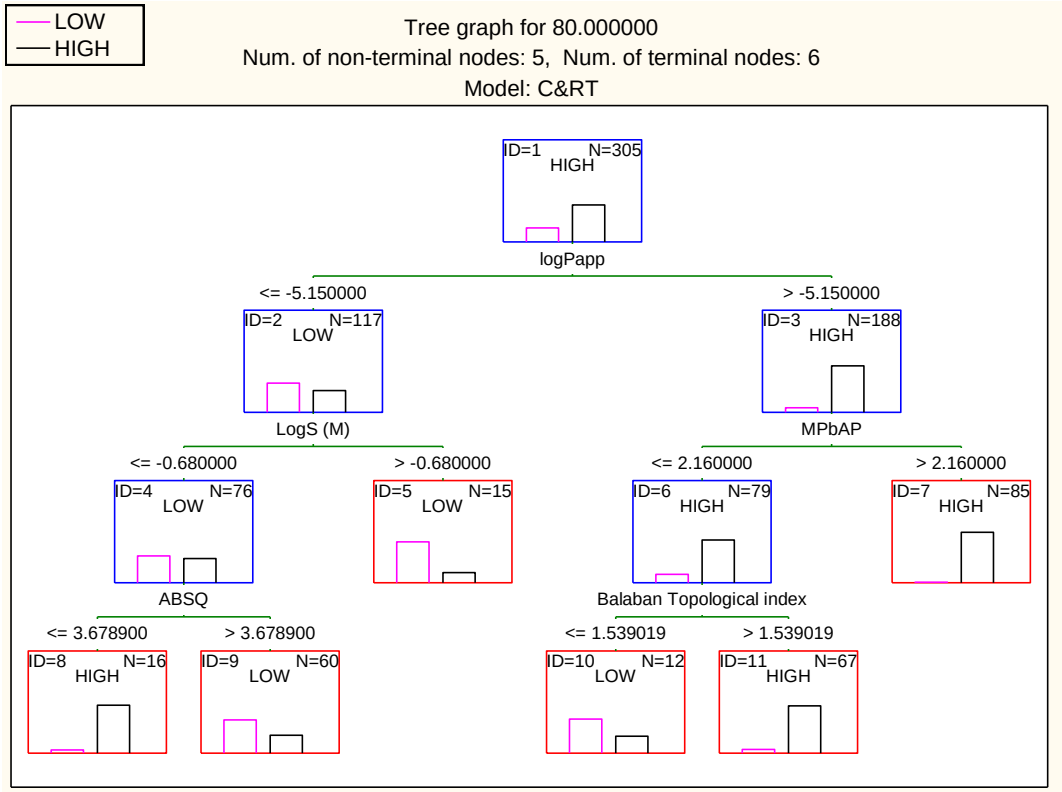

Figure A3. 15. Model 8; C&RT permeability experimental solubility (logS M) and melting point based absorption potential (MPbAP) model when higher misclassification costs of two were applied to the high MPbAP node only

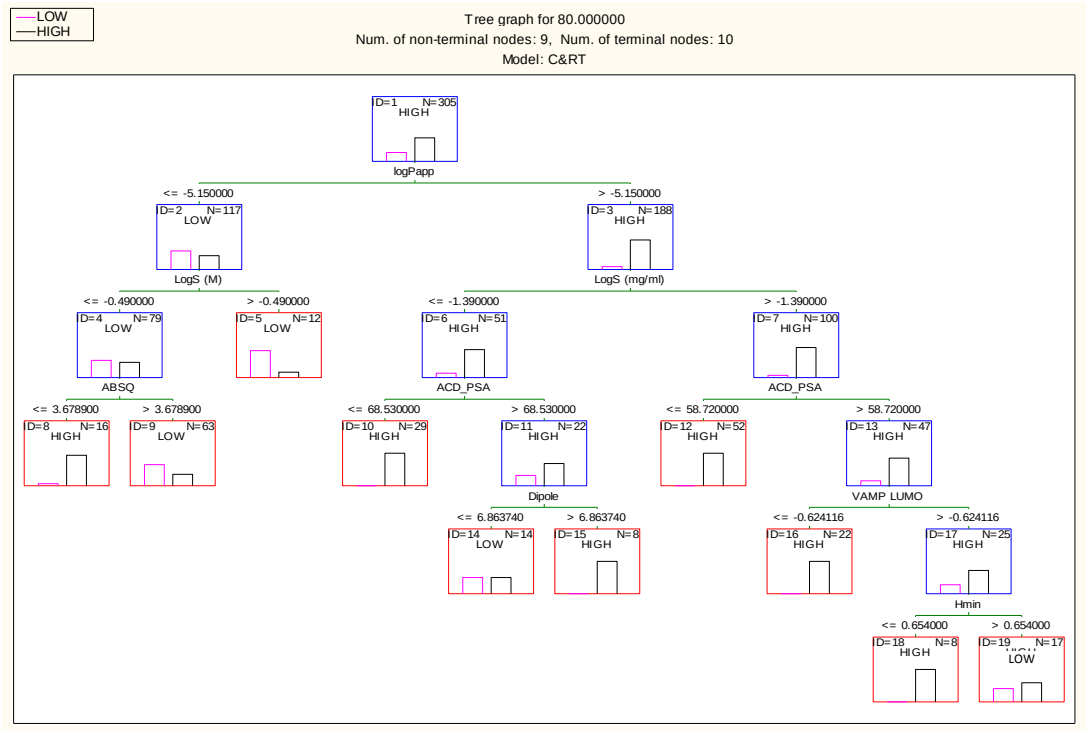

Figure A3. 16. Model 9; C&RT permeability and experimental solubility (logS in M and mg/mL) model when higher misclassification costs of two and ten to reduce false positives were applied to high logS (mg/mL) node and low logS (M) solubility node respectively

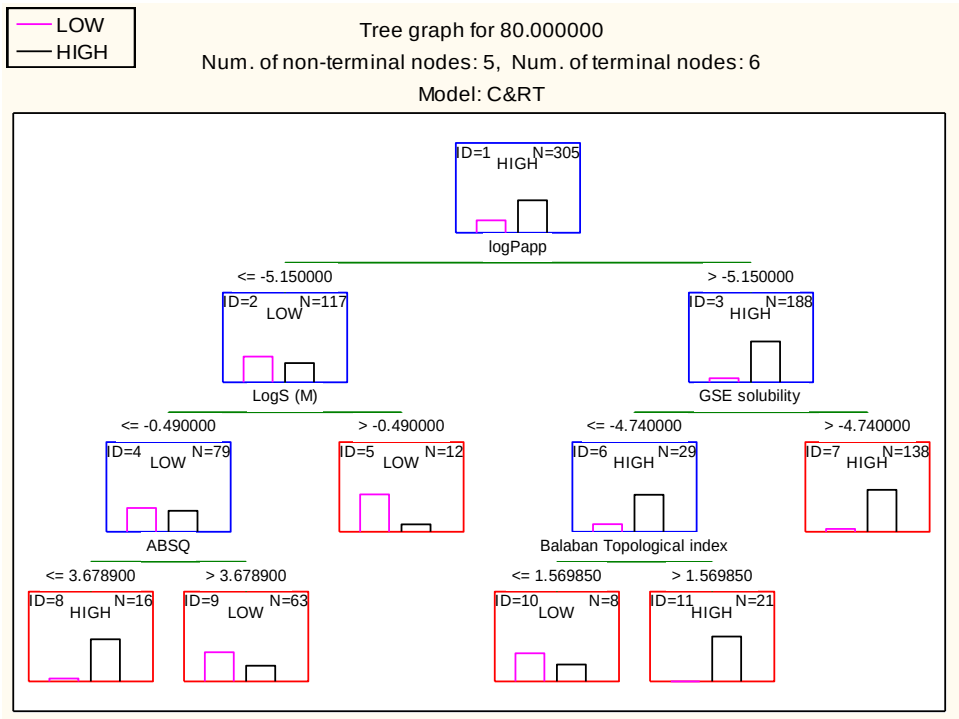

Figure A3. 17. Model 10; C&RT permeability, experimental solubility (logS M) and predicted solubility (GSE) model when higher misclassification costs of two to reduce false positives were applied to high GSE solubility node only

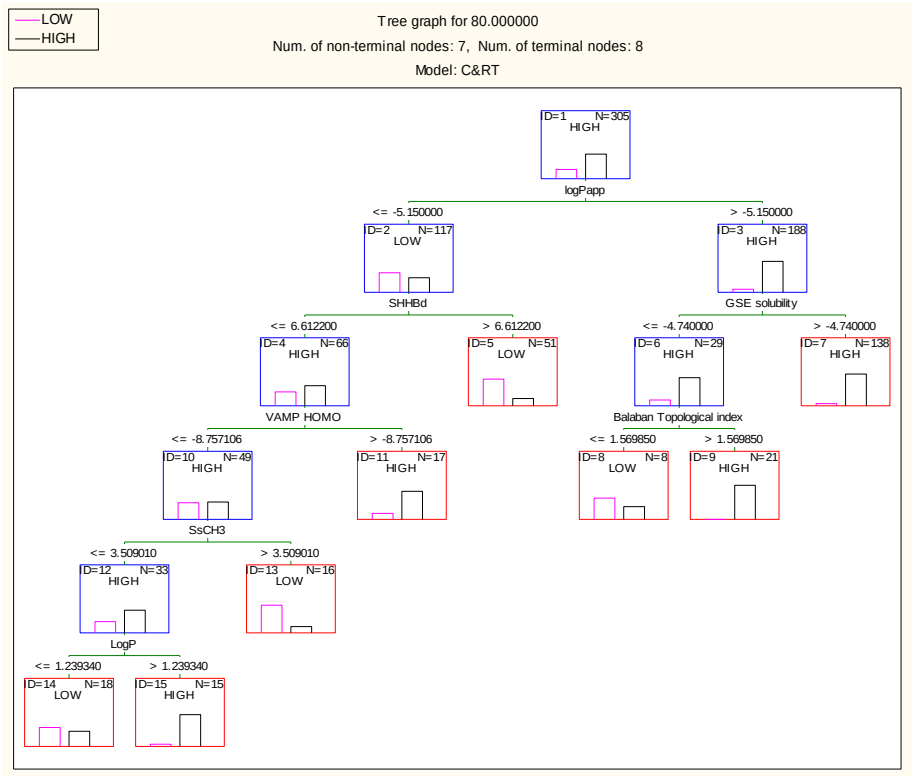

Figure A3. 18. Model 11; C&RT permeability and predicted solubility (GSE) model when higher misclassification costs of two to reduce false positives were applied to high GSE solubility node only

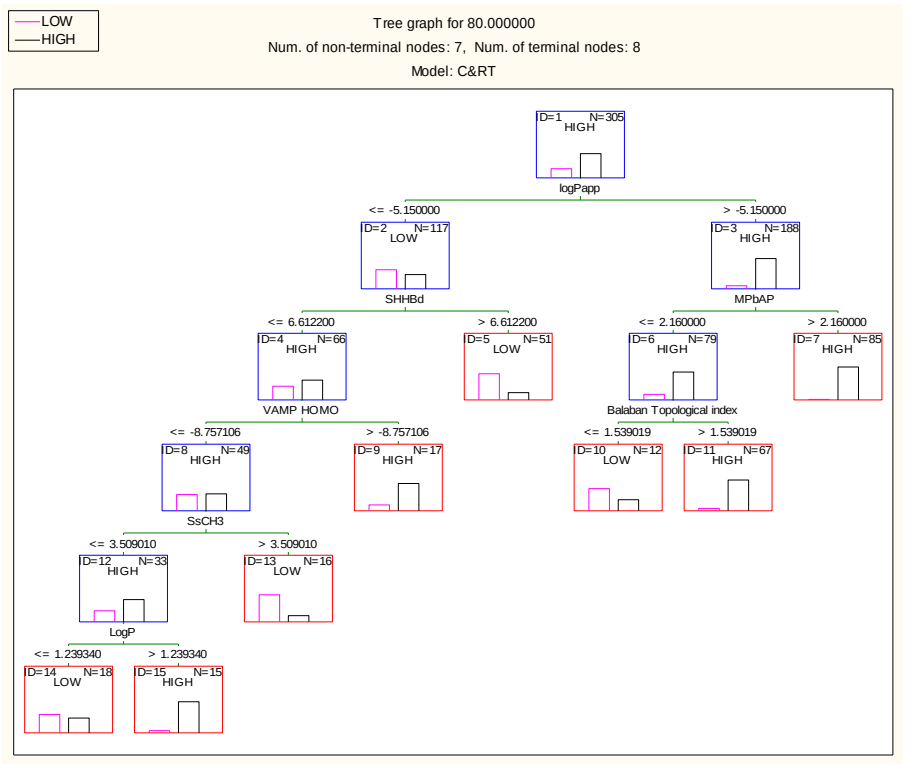

Figure A3. 19. Model 12; C&RT permeability and melting point based absorption potential (MPbAP) model when higher misclassification costs of two to reduce false positives were applied to high MPbAP solubility node

Appendix 4: Supporting Information for Chapter 10

Table A4. 1. The top 20 molecular descriptors selected by variable importance using random forest for solubility class

Descriptor	*Description*
LogD(2)	Apparent distribution coefficient at pH 2 calculated by ACD
LogD(6.5)	Apparent distribution coefficient at pH 6.5 calculated by ACD
LogD(5.5)	Apparent distribution coefficient at pH 5.5 calculated by ACD
PEOE_VSA_POL	Total polar van der Waals surface area using calculated partial charges by PEOE
PEOE_VSA_HYD	Total hydrophobic van der Waals surface area using calculated partial charges by PEOE
FIBpH6.5	Fraction of drug ionised as bases as pH 6.5
BCUT_PEOE_0	The BCUT descriptors are calculated from the eigenvalues of a modified adjacency matrix using calculated partial charges by PEOE
LogD(7.4)	Apparent distribution coefficient at pH 7.4 calculated by ACD
PEOE_VSA_FPOS	Fractional positive van der Waals surface area using calculated partial charges by PEOE
FIBpH5.5	Fraction of drug ionised as bases as pH 5.5
BCUT_SLOGP_2	The BCUT descriptors are calculated from the eigenvalues of a modified adjacency matrix using atomic contribution to logP instead of partial charge.
vsurf_Wp4	Polar volume
b_single	Number of single bonds (including implicit hydrogens).
FIBpH2	Fraction of drug ionised as bases as pH 2
VDistEq	VDistEq is related to the size and shape of a molecule.
GCUT_PEOE_3	The GCUT descriptors are calculated from the eigenvalues of a modified graph distance adjacency matrix using calculated partial charges by PEOE
PEOE_VSA_FHYD	Fractional hydrophobic van der Waals surface area
FCASA-	Fractional negative charge weighted surface area
rgyr	Radius of gyration
MaxHp	Largest positive charge on a hydrogen atom

Table A4. 2. The top 20 molecular descriptors selected by variable importance using random forest for permeability class

Descriptor	*Description*
LogD(6.5)	Apparent distribution coefficient at pH 6.5 calculated by ACD
vsurf_CW4	Capacity factor
LogD(5.5)	Apparent distribution coefficient at pH 5.5 calculated by ACD
vsurf_HB1	H-bond donor capacity
LogD(10)	Apparent distribution coefficient at pH 10 calculated by ACD
xv2	Valence chi 2 index
vsurf_W3	Hydrophilic volume
chi1_C	Carbon connectivity index (order 1)
FIBpH7.4	Fraction of drug ionised as base as pH 7.4
vsurf_HL2	Hydrophilic-Lipophilic balance
PEOE_RPC-	Relative partial charge calculated using PEOE
LogD(7.4)	Apparent distribution coefficient at pH 7.4 calculated by ACD
vsurf_Wp2	Polar Volume
vsurf_HL1	Hydrophilic-Lipophilic balance
PSA	Polar Surface area
vsurf_W2	Hydrophilic volume
FIBpH6.5	Fraction of drug ionised as bases as pH 6.5
GCUT_PEOE_0	The GCUT descriptors are calculated from the eigenvalues of a modified graph distance adjacency matrix using calculated partial charges by PEOE
vsurf_Wp3	Polar Volume
FIAB_6.5	Fraction of drug ionised as base multiplied by fraction ionised as anion calculated at pH 6.5

Table A4. 3. Experimental and literature Biopharmaceutics Classification System (BCS) class comparison for external validation set (n=127) in chapter 10

#	Compound Name	Solubility (mg/mL)	Log Papp	Permeability Reference	Exp BCS class	Lit BCS class	Literature BCS References	Differences between literature and experimental BCS class assignment
1	Praziquantel	0.4	-4.357	(Gonzalez-Esquivel *et al*., 2005)	1	2	(WHO, 2006 (accessed December 19, 2013))	Large maximum dose causes a large dose number and therefore low solubility definition according to FDA guidelines
2	Verapamil	0.7205	-4.438	(Varma *et al*., 2005, Polli *et al*., 2001, Skolnik *et al*., 2010)	1	2	(WHO, 2006 (accessed December 19, 2013))	Large maximum dose causes a large dose number and therefore low solubility definition according to FDA guidelines
3	Acetazolamide	0.7112	-6.678	(Granero *et al*., 2008, Crowe and Teoh, 2006)	3	2/4	(WHO, 2006 (accessed December 19, 2013))	Large maximum dose causes a large dose number and therefore low solubility definition according to FDA guidelines
4	Gemfibrozil	>0.5	-4.407	(Huang *et al*., 2010)	1	2	(Bergman *et al*., 2010)	Large maximum dose causes a large dose number and therefore low solubility definition according to FDA guidelines
5	Biperiden	0.01 (PI)	-4.081	(Abalos *et al*., 2012)	2	1/3	(WHO, 2006 (accessed December 19, 2013), Cao *et al*., 2012)	The WHO BCS class assignment was on the basis on insufficient permeability data based on permeability data can now be assigned either 1/2 due to differences in free base and hydrochloride salt formulation
6	Paricalcitol	0.01 (PI)	-5.103	(Palaparthy *et al*., 2007)	2	4	http://www.accessdata.fda.gov/drugsatfda_docs/nda/2005/021606s000_ClinPharmR.pdf (Accessed 5 Jan 2014)	Permeability is close to cut off threshold so with variability of caco-2 data could therefore be classed as poorly permeable, also HIA is less than 90% could explain why BCS assignment is wrong if based on *in vivo* data
7	Doxepin	0.03157	-4.266	(Varma *et al*., 2005)	2	1	http://www.accessdata.fda.gov/drugsatfda_docs/nda/2010/022036Orig1s000ClinPharmR.pdf (accessed 18 December 2013)	Based on HCl salt assigned BCS class 1, solubility value based on free base
8	Vardenafil	0.11	-5.252	(Choi and Song, 2012)	4	2	http://www.ema.europa.eu/docs/en_GB/document_library/EPAR__Assessment_Report__Variation/human/000475/WC500097073.pdf (Accessed 5 Jan 2014) (Choi and Song, 2012)	Low permeability (in Caco-2) due to efflux effects by multiple transporters where as in MDCK-MDR1 higher permeability, differences between *in vitro*/*in vivo*. Based on %HIA of 90% this compound would be classed as highly-absorbed (Choi and Song, 2012)

#	Compound Name	Solubility (mg/mL)	Log Papp	Permeability Reference	Exp BCS class	Lit BCS class	Literature BCS References	Differences between literature and experimental BCS class assignment
9	Pregabalin	32.1	-5.928	(Jezyk *et al*., 1999)	3	1	(Cook *et al*., 2008)	Substrate for a Ltype uptake transporter therefore differences could be due to IVIV transporter expression. The literature states that caco-2 is not a suitable model for highlighting the effect of transporter mediated transport for this compound. (Jezyk *et al*., 1999, Su *et al*., 2005)
10	Aspirin	4.46	-5.319	(Irvine *et al*., 1999)	3	1	(Dressman *et al*., 2012, Lindenberg *et al*., 2004)	A weak acid and is absorbed in lower part of the intestine (pH 5.4), permeability carried out at pH 7.4 therefore possibly explains low experimental permeability (Levitt, 2013)
11	Cefixime	1 (SS)	-5.921	(Balimane *et al*., 2007)	3	2/4	(WHO, 2006 (accessed December 19, 2013))	pH dependent solubility and literature states ionized at physiological pH therefore solubility could be lower as suspension pH is low Based on solubility definition according to FDA across a wide pH range could explain difference in solubility. Evidence suggests this compound is substrate for a influx carrier mediated transporter in the small intestine therefore difference could also IVIV differences (Tsuji *et al*., 1987, Wenzel *et al*., 2002) (http://products.sanofi.ca/en/suprax.pdf (Accessed 5Jan2014)).
12	Abacavir	1.68E-06	-5.463	(Shaik *et al*., 2007)	4	3	(WHO, 2006 (accessed December 19, 2013)) http://www.accessdata.fda.gov/drugsatfda_docs/label/2013/020977s026,020978s030lbl.pdf (Accessed 7 Jan 2014)	Experimental solubility value from free base of compound. However sulfate of compound is highly soluble (77mg/ml) and this could be why compound is classed as BCS class 3. In addition this compound is a substrate for pgp and this could explain the low permeability in the MDCK-MDR1 cell line http://www.accessdata.fda.gov/drugsatfda_docs/label/2013/020977s026,020978s030lbl.pdf (Accessed 6Jan 2014) (Shaik *et al*., 2007)
13	Venlafaxine	<0.1	-4.666	(Wager *et al*., 2010, Feng *et al*., 2008)	2	1	(Ramirez *et al*., 2010)	Martindale states FS for hydrochloride salt, free base much lower solubility compared with marketed product (hydrochloride salt 572 mg/ml). Therefore BCS differences due to formulation (Martindale, 2009)

#	Compound Name	Solubility (mg/mL)	Log Papp	Permeability Reference	Exp BCS class	Lit BCS class	Literature BCS References	Differences between literature and experimental BCS class assignment
14	Acetaminophen	14.475	-4.448	(Sherer *et al*., 2012, Chen *et al*., 2005)	1	3	(WHO, 2006 (accessed December 19, 2013), Kalantzi *et al*., 2006)	Based on other *in vitro*/in situ tests compounds classed as poorly permeable in addition literature stated that *in vivo* absorption could be higher than 80% therefore differences between *in vitro*/*in vivo* and the thresholds used and the uncertainty surrounding this compounds assignment (Kalantzi *et al*., 2006)
15	Sertraline hydrochloride	4.24	-5.629	(Feng *et al*., 2008, Ingels, 2004)	3	1	http://www.accessdata.fda.gov/drugsatfda_docs/label/2013/019839s079,020990s038lbl.pdf (Accessed 7 Jan 2014) (Ramirez *et al*., 2010)	Compound is stated to be absorbed but very slowly therefore this could explain difference between classes; in addition this compound can inhibit certain transporters. Literature indicated compound (as hydrochloride salt) is slightly soluble. Inhibitor of pgp, PAT1(non competitive). Evidence suggests a pgp substrate and inhibitor (Nielsen *et al*., 2013, Wang *et al*., 2008)
16	Paroxetine	6.213	-5.456	(Wager *et al*., 2010, Feng *et al*., 2008)	3	1	http://www.hma.eu/fileadmin/dateien/pipar/dk233paroxetinhexal/mod5_par_dk233_03_04_paroxetin_hexal_20060620.pdf (Accessed 13 Jan 2014)	Strong inhibitor of pgp and SERT, Conflicting literature stating whether or not this compound is a pgp substrate however efflux ratios collected from the literature seem to indicate it is a substrate and also a inhibitor. (Loscher and Potschka, 2005, Wager *et al*., 2010, Maines *et al*., 2005, Kikuchi *et al*., 2013)
17	Mirtazapine	0.01 (PI)	-4.587	(Wager *et al*., 2010)	2	1/2	(Ramirez *et al*., 2010) http://www.pfizer.ca/en/our_products/products/monograph/320	Indicated pH dependent solubility, high pH becomes less soluble. Could be highly soluble if used label information (slightly soluble 1mg/ml) compared with Martindale (PI 0.01mg/ml) which would be poorly soluble- we took solubility for the worse possible scenario therefore 0.01mg/ml was used. http://www.mhra.gov.uk/home/groups/l-unit1/documents/websiteresources/con2032430.pdf (accessed 15 Jan 2014)(Martindale, 2009)
18	Isoniazid	122.8	-4.664	(Ranaldi *et al*., 1992)	1	1/3	(WHO, 2006 (accessed December 19, 2013))	
19	Ethosuximide	190	-4.572	(Feng *et al*., 2008)	1	1/3	(WHO, 2006 (accessed December 19, 2013))	

#	Compound Name	Solubility (mg/mL)	Log Papp	Permeability Reference	Exp BCS class	Lit BCS class	Literature BCS References	Differences between literature and experimental BCS class assignment
20	Clomipramine	100 (FS)	-4.433	(Varma *et al*., 2005)	1	1/3	(WHO, 2006 (accessed December 19, 2013))	
21	Furosemide	0.0062	-6.235	(Varma *et al*., 2012, Irvine *et al*., 1999, Rege *et al*., 2001)	4	2/4	(WHO, 2006 (accessed December 19, 2013))	
22	Amoxicillin	4	-6.475	(Sherer *et al*., 2012, Irvine *et al*., 1999)	3	1/3	(WHO, 2006 (accessed December 19, 2013))	
23	Antipyrine	637	-4.312	(Varma *et al*., 2012, Varma *et al*., 2005, Gres *et al*., 1998, Hilgendorf *et al*., 2000, Yamashita *et al*., 2000, Garberg *et al*., 2005, Jung *et al*., 2006)	1	1/2	http://www.fda.gov/downloads/Drugs/.../Guidances/ucm070246.pdf (accessed 19 December 2013)	
24	Rilpivirine	0.01 (PI)	-4.924	http://www.accessdata.fda.gov/drugsatfda_docs/nda/2011/202022Orig1s000ClinPharmR.pdf	2	2	(Mathias *et al*., 2012)	
25	Cinacalcet	<0.001	-5.237	http://www.accessdata.fda.gov/drugsatfda_docs/nda/2004/21-688.pdf_Sensipar_BioPharmr.pdf	4	4	http://www.accessdata.fda.gov/drugsatfda_docs/nda/2004/21-688.pdf_Sensipar_BioPharmr.pdf (Accessed 7 Jan 2014)	
26	Metronidazole	10.3	-4.735	(Varma *et al*., 2012)	1	1	(WHO, 2006 (accessed December 19, 2013))	
27	Voriconazole	0.1 (VSS)	-4.551	(Damle *et al*., 2011)	2	2	http://www.accessdata.fda.gov/drugsatfda_docs/nda/2003/021630s000_Vfend_ClinPharm.pdf (Accessed 19 December 2013)	
28	Ethambutol	649	-6.097	(Varma *et al*., 2012)	3	3	(WHO, 2006 (accessed December 19, 2013))	
29	Pyridostigmine	1000 (VS)	-5.991	(Varma *et al*., 2005)	3	3	(WHO, 2006 (accessed December 19, 2013))	

#	Compound Name	Solubility (mg/mL)	Log Papp	Permeability Reference	Exp BCS class	Lit BCS class	Literature BCS References	Differences between literature and experimental BCS class assignment
30	Ascorbic_Acid	249	-5.143	(Lu *et al*., 2010, Luo *et al*., 2008)	1	1	(WHO, 2006 (accessed December 19, 2013))	
31	Propranolol	360	-4.480	(Varma *et al*., 2005, Skolnik *et al*., 2010, Mahar Doan *et al*., 2002)	1	1	(WHO, 2006 (accessed December 19, 2013))	
32	Ranitidine	550	-6.258	(Varma *et al*., 2012, Rege *et al*., 2001)	3	3	(WHO, 2006 (accessed December 19, 2013))	
33	Atenolol	26.5	-6.097	(Jung *et al*., 2006, Artursson and Karlsson, 1991, Irvine *et al*., 1999, Varma *et al*., 2012, Varma *et al*., 2005, Yamashita *et al*., 2000, Hilgendorf *et al*., 2000, Stenberg *et al*., 2001, Nordqvist *et al*., 2004, Yazdanian *et al*., 1998, Aungst *et al*., 2000, Collett *et al*., 1996)	3	3	(WHO, 2006 (accessed December 19, 2013))	
34	Dapsone	0.15	-4.957	(Monteiro *et al*., 2012)	2	2	(WHO, 2006 (accessed December 19, 2013))	
35	Diethylcarbamazine	1000 (VS)	-4.881	(Kogan *et al*., 2008)	1	1	(Lindenberg *et al*., 2004)	
36	Pramipexole	>10	-4.893	(Wager *et al*., 2010)	1	1	http://www.accessdata.fda.gov/drugsatfda_docs/nda/2010/022421s000chemr.pdf (accessed 19 December 2013)	
37	Theophylline	7.55	-4.505	(Sherer *et al*., 2012, Varma *et al*., 2012)	1	1	(Lindenberg *et al*., 2004)	
38	Glimepiride	>0.0736	-4.517	(Frick *et al*., 1998)	2	2	(Taupitz *et al*., 2013, Nagpal *et al*., 2012)	
39	Perphenazine	0.02828	-4.516	(Varma *et al*., 2005)	2	2	(Baboota *et al*., 2013)	

#	Compound Name	Solubility (mg/mL)	Log Papp	Permeability Reference	Exp BCS class	Lit BCS class	Literature BCS References	Differences between literature and experimental BCS class assignment
40	Enzalutamide	0.002	-4.509	http://www.accessdata.fda.gov/drugsatfda_docs/nda/2012/203415Orig1s000ClinPharmR.pdf	2	2	http://www.accessdata.fda.gov/drugsatfda_docs/nda/2012/203415Orig1s000ClinPharmR.pdf (Accessed 19Dec2013)	
41	Clonazepam	0.1	-4.507	(Wager *et al*., 2010)	2	2	(Subramanian *et al*., 2012, Nainar *et al*., 2012)	
42	Loratadine	0.01	-4.429	(Varma *et al*., 2005, Khan *et al*., 2004)	2	2	(Khan *et al*., 2004, Ramirez *et al*., 2010).	
43	Carbamazepine	0.15	-4.341	(Varma *et al*., 2012, Varma *et al*., 2005, Skolnik *et al*., 2010)	2	2	(Kovacevic *et al*., 2009)	
44	Lercanidipine	0.005	-7.301	US Patent 2006/0073200 A1, Serial number EP1807059 A1	4	4	(Granero *et al*., 2010)	
45	Trospium	500	-6.830	(Langguth *et al*., 1997)	3	3	(Radwan *et al*., 2012)	
46	Crizotinib	<0.1	-5.170	http://www.accessdata.fda.gov/drugsatfda_docs/nda/2011/202570Orig1s000ClinPharmR.pdf	4	4	http://www.ema.europa.eu/docs/en_GB/document_library/EPAR__Public_assessment_report/human/002489/WC500134761.pdf (Accessed 19 December 2013)	
47	Caffeine	21.6	-4.420	(Feng *et al*., 2008, Varma *et al*., 2012, Yazdanian *et al*., 1998, Garberg *et al*., 2005, Jung *et al*., 2006, Gres *et al*., 1998, Yee, 1997, Chong *et al*., 1996)	1	1	(Wu and Benet, 2005, Smetanova *et al*., 2009)	
48	Memantine	38.6	-4.363	(Beconi *et al*., 2011)	1	1	http://www.accessdata.fda.gov/drugsatfda_docs/nda/2003/21487_namenda_bioeqr_p1.pdf (Accessed 19 December 2013) http://www.accessdata.fda.gov/drugsatfda_docs/nda/2005/021627s000_namenda_clinpharmr.pdf (Accessed 13 January 2014)	

#	Compound Name	Solubility (mg/mL)	Log Papp	Permeability Reference	Exp BCS class	Lit BCS class	Literature BCS References	Differences between literature and experimental BCS class assignment
49	Gliclazide	0.025	-4.602	(Smetanova *et al*., 2009)	2	2	(Benet *et al*., 2011, Grbic *et al*., 2011)	
50	Cimetidine	9.3	-5.911	(Varma *et al*., 2012, Varma *et al*., 2005, Skolnik *et al*., 2010)	3	3	(Jantratid *et al*., 2006)	
51	Clopidogrel bisulfate	0.01	-4.896	(Ki *et al*., 2008)	2	2	(Ramirez *et al*., 2010, Lassoued *et al*., 2011)	
52	Nalidixic_Acid	0.1	-3.844	(Ranaldi *et al*., 1992)	2	2	(Katdare and Chaubal, 2006)	
53	Glyburide	0.031	-4.614	(Varma *et al*., 2012, Zerrouk *et al*., 2006)	2	2	(Wei and Lobenberg, 2006)	
54	Quinine	0.5497	-4.498	(Crivori *et al*., 2006)	1	1	(WHO, 2006 (accessed December 19, 2013), Lindenberg *et al*., 2004)	
55	Metoprolol	43	-4.677	(Varma *et al*., 2012, Varma *et al*., 2005, Skolnik *et al*., 2010)	1	1	(Tsume and Amidon, 2010)	
56	Ruxolitinib	2.7	-4.668	http://www.accessdata.fda.gov/drugsatfda_docs/nda/2011/202192Orig1s000ClinPharmR.pdf	1	1	http://ec.europa.eu/health/documents/community-register/2012/20120823123254/anx_123254_en.pdf (Accessed 2 January 2014)	
57	Itraconazole	0.004	-4.824	(Varma *et al*., 2012)	2	2	http://www.accessdata.fda.gov/drugsatfda_docs/nda/2010/022484Orig1s000ClinPharmR.pdf (Accessed 2 Jan 2014)	
58	Nimodipine	0.003055	-4.812	(Wager *et al*., 2010)	2	2	(Papageorgiou *et al*., 2009)	
59	Galantamine Hydrobromine	31	-4.807	(Wager *et al*., 2010)	1	1	http://www.accessdata.fda.gov/drugsatfda_docs/nda/2001/21-224_REMINYL_biopharmr.pdf (accessed 13 Jan 2014)	
60	Norfloxacin	0.19	-5.627	(Skolnik *et al*., 2010)	4	4	(Breda *et al*., 2009)	
61	Phenylbutazone	0.034	-4.998	(Khan *et al*., 2011)	2	2	(Bhakay *et al*., 2013)	
62	Amphotericin_B	0.003	-6.921	(Skolnik *et al*., 2010)	4	4	(Wu and Benet, 2005, Yanez *et al*., 2011)	
63	Milnacipran	100 (FS)	-5.051	(Dyck *et al*., 2008)	1	1	http://www.accessdata.fda.gov/drugsatfda_docs/nda/2009/022256s000_CDTLMemo.pdf(accessed 19 December 2019)	

#	Compound Name	Solubility (mg/mL)	Log Papp	Permeability Reference	Exp BCS class	Lit BCS class	Literature BCS References	Differences between literature and experimental BCS class assignment
64	Quetiapine	0.4	-4.662	(Wager *et al*., 2010)	2	2	http://www.accessdata.fda.gov/drugsatfda_docs/nda/2007/022047Orig1s000ChemR.pdf (Accessed 2 Jan 2014)	
65	Hydrocodone	33 (S)	-4.606	(Wager *et al*., 2010, Feng *et al*., 2008)	1	1	(Hemmingsen *et al*., 2011)	
66	Olanzapine	0.01 (PI)	-4.726	(Wager *et al*., 2010)	2	2	(Thakuria and Nangia, 2011, Dixit *et al*., 2011)	
67	Rivaroxaban	0.01 (PI)	-5.092	(Gnoth *et al*., 2011)	2	2	http://www.accessdata.fda.gov/drugsatfda_docs/nda/2011/202439Orig1s000ClinPharmR.pdf (Accessed 5 Jan 2014)	
68	Citalopram Hydrobromide	100 (FS)	-4.814	(Wager *et al*., 2010, Feng *et al*., 2008) http://www.acrossbarriers.eu/uploads/media/FCT02-I-0305_BCS.pdf	1	1	(Ramirez *et al*., 2010)	
69	Cefotaxime	100 (FS)	-5.807	(Raeissi *et al*., 1999)	3	3	(Sharma *et al*., 2005)	
70	Methylphenidate	100 (FS)	-4.577	(Wager *et al*., 2010, Feng *et al*., 2008)	1	1	http://www.ema.europa.eu/docs/en_GB/document_library/Referrals_document/Methylphenidate_Hexal/WC500156885.pdf (Accessed 13 Jan 2014)	
71	Indomethacin	0.002	-4.377	(Varma *et al*., 2012, Varma *et al*., 2005, Sherer *et al*., 2012)	2	2	(ElShaer *et al*., 2011, Clarysse *et al*., 2009)	
72	Tiagabine	10 (SPS)	-4.684	(Wager *et al*., 2010)	1			
73	Ribavirin	142	-6.745	(Li *et al*., 2006)	3			
74	Atazanavir sulfate	4.5	-5.921	(Kis *et al*., 2013)	3			
75	Mexiletine	100 (FS)	-3.916	(Catalano *et al*., 2012)	1			
76	Mefloquine	1.806	-5.027	(Milner *et al*., 2010)	1			
77	Acamprosate	100 (FS)	-5.986	(Zornoza *et al*., 2004)	3			
78	Selegiline	100 (FS)	-4.153	(Varma *et al*., 2005)	1			
79	Cefamandole Nafate	100 (FS)	-5.650	(Raeissi *et al*., 1999)	3			

#	Compound Name	Solubility (mg/mL)	Log Papp	Permeability Reference	Exp BCS class	Lit BCS class	Literature BCS References	Differences between literature and experimental BCS class assignment
80	Doxapram	10 (SPS)	-4.346	(Varma *et al*., 2005)	1			
81	Methysergide	1 (SS)	-4.759	(Varma *et al*., 2005)	1			
82	Dofetilide	0.1 (VSS)	-4.804	(Singleton *et al*., 2007)	2			
83	Zonisamide	0.8	-4.536	(Wager *et al*., 2010)	1			
84	Ramelteon	0.21	-4.476	(Wager *et al*., 2010)	1			
85	Trimipramine	0.0048	-4.393	(Varma *et al*., 2005)	2			
86	Tacrine	0.92	-4.583	(Varma *et al*., 2005, Kuo *et al*., 2012)	1			
87	Darifenacin hydrobromide	6.03	-4.721	(Skerjanec, 2006)	1			
88	Flumazenil	0.1 (VSS)	-4.224	(Varma *et al*., 2005)	2			
89	Piperacillin	100 (FS)	-7.488	(Violette *et al*., 2008)	3			
90	Naltrexone	100 (FS)	-4.558	(Varma *et al*., 2005, Kanaan *et al*., 2009)	1			
91	Cefadroxil	1 (SS)	-5.056	(Raeissi *et al*., 1999)	1			
92	Dolasetron	100 (FS)	-4.889	(Dow *et al*., 1996)	1			
93	Reboxetine	250	-4.914	(Wager *et al*., 2010)	1			
94	Oseltamivir phosphate	>500	-5.209	(Oo *et al*., 2003)	3			
95	Rivastigmine	1000 (VS)	-4.564	(Wager *et al*., 2010)	1			
96	Riluzole	0.1 (VSS)	-4.517	(Wager *et al*., 2010)	2			
97	Zaleplon	0.01 (PI)	-4.431	(Wager *et al*., 2010)	2			
98	UK-294,315	0.74	-4.951	(Harrison *et al*., 2004)	1			
99	Nalbuphine hydrochloride	35.5	-4.854	(Varma *et al*., 2005)	1			

#	Compound Name	Solubility (mg/mL)	Log Papp	Permeability Reference	Exp BCS class	Lit BCS class	Literature BCS References	Differences between literature and experimental BCS class assignment
100	Prazosin	0.0032	-5.019	(Di *et al*., 2011, Skolnik *et al*., 2010, Varma *et al*., 2005)	2			
101	Minocycline	52	-5.310	(Varma *et al*., 2012)	3			
102	Ropinirole Hydrochloride	133	-4.570	(Wager *et al*., 2010)	1			
103	Cephalothin	0.457	-5.349	(Raeissi *et al*., 1999)	3			
104	Norethindrone	0.00633	-4.770	(Faassen *et al*., 2003, Kim and Benet, 2004)	2			
105	Propoxyphene	0.0037	-4.660	(Feng *et al*., 2008)	2			
106	Flurazepam	500	-4.152	(Varma *et al*., 2005)	1			
107	Cefaclor	10	-5.824	(Balimane *et al*., 2007)	3			
108	Terbutaline	90	-5.772	(Skolnik *et al*., 2010, Irvine *et al*., 1999)	3			
109	Maprotiline	0.000833	-4.342	(Varma *et al*., 2005)	2			
110	Oxycodone	142.9	-4.638	(Wager *et al*., 2010, Hassan *et al*., 2007)	1			
111	Pheniramine	11.05	-4.754	(Varma *et al*., 2005, Marasanapalle *et al*., 2009)	1			
112	Cilostazol	0.000101	-4.699	(Young *et al*., 2006)	2			
113	Methimazole	~200	-4.319	(Skold *et al*., 2006)	1			
114	Famciclovir	20	-5.148	(Varma *et al*., 2005)	1			
115	Nitrazepam	0.043	-4.410	(Varma *et al*., 2005)	2			
116	Hydroxyzine	100 (FS)	-5.035	(Feng *et al*., 2008, Laitinen *et al*., 2003)	1			
117	Oxybutynin	0.8	-4.456	(Callegari *et al*., 2011)	1			
118	Gemifloxacin	0.35	-4.921	(Jin *et al*., 2013)	1			

#	Compound Name	Solubility (mg/mL)	Log Papp	Permeability Reference	Exp BCS class	Lit BCS class	Literature BCS References	Differences between literature and experimental BCS class assignment
119	Indacaterol maleate	0.1 (VSS)	-4.703	http://www.accessdata.fda.gov/drugsatfda_docs/nda/2011/022383Orig1s000ClinPharmR.pdf (accessed 5 Jan 2014)	2			
120	Vildagliptin	>50	-5.824	(He *et al*., 2009)	3			
121	Atomoxetine	27.8	-4.777	(Wager *et al*., 2010)	1			
122	Fluvoxamine	14.869	-4.499	(Varma *et al*., 2005)	1			
123	Pergolide Mesilate	1 (SS)	-5.013	(Wager *et al*., 2010)	1			
124	Procyclidine	0.001055	-4.153	(Varma *et al*., 2005)	2			
125	Nortriptyline	0.025	-4.472	(Varma *et al*., 2005)	2			
126	Methyl phenobarbital	0.15	-4.387	(Behrens *et al*., 2001)	2			
127	Diphenoxylate	0.8	-6.699	(Crowe and Wong, 2003)	3			

Exp; experimental BCS class using collected data and thresholds defined in chapter 10; Lit: Literature cited BCS class. FR: freely soluble; PI: practically insoluble; SS: slightly soluble; SPS: sparingly soluble; VS: very soluble; VSS: very slightly soluble

References for Table A4.3

ABALOS, I. S., RODRIGUEZ, Y. I., LOZANO, V., CERESETO, M., MUSSINI, M. V., SPINETTO, M. E., CHIALE, C. & PESCE, G. 2012. Transepithelial transport of biperiden hydrochloride in Caco-2 cell monolayers. *Environmental Toxicology and Pharmacology,* 34**,** 223-7.

ARTURSSON, P. & KARLSSON, J. 1991. Correlation between oral-drug absorption in humans and apparent drug perrmeability coefficients in human intestinal epithelial (Caco-2) cells. *Biochemical and Biophysical Research Communications,* 175**,** 880-885.

AUNGST, B. J., NGUYEN, N. H., BULGARELLI, J. P. & OATES-LENZ, K. 2000. The influence of donor and reservoir additives on Caco-2 permeability and secretory transport of HIV protease inhibitors and other lipophilic compounds. *Pharmaceutical Research,* 17**,** 1175-80.

BABOOTA, S., ABDULLAH, M., GULAM, S., JASJEET K. & ALI, J. 2013. Mechanistic approach for the development of ultrafine oil-water emulsions using monoglyceride and blends of medium and long chain triglycerides: enhancement of the solubility and bioavailability of Perphenazine. *Journal of Excipients & Food Chemicals,* 4**,** 12.

BALIMANE, P. V., CHONG, S., PATEL, K., QUAN, Y., TIMOSZYK, J., HAN, Y. H., WANG, B., VIG, B. & FARIA, T. N. 2007. Peptide transporter substrate identification during permeability screening in drug discovery: Comparison of transfected MDCK-hPepT1 cells to Caco-2 cells. *Archives of Pharmacal Research,* 30**,** 507-518.

BECONI, M. G., HOWLAND, D., PARK, L., LYONS, K., GIULIANO, J., DOMINGUEZ, C., MUNOZ-SANJUAN, I. & PACIFICI, R. 2011. Pharmacokinetics of memantine in rats and mice. *PLoS Curr,* 15.

BEHRENS, I., STENBERG, P., ARTURSSON, P. & KISSEL, T. 2001. Transport of lipophilic drug molecules in a new mucus-secreting cell culture model based on HT29-MTX cells. *Pharmaceutical Research,* 18**,** 1138-45.

BENET, L. Z., BROCCATELLI, F. & OPREA, T. I. 2011. BDDCS applied to over 900 drugs. *The AAPS Journal,* 13**,** 519-47.

BERGMAN, E., MATSSON, E. M., HEDELAND, M., BONDESSON, U., KNUTSON, L. & LENNERNAS, H. 2010. Effect of a single gemfibrozil dose on the pharmacokinetics of rosuvastatin in bile and plasma in healthy volunteers. *Journal of Clinical Pharmacology,* 50**,** 1039-49.

BHAKAY, A., DAVÉ, R. & BILGILI, E. 2013. Recovery of BCS Class II drugs during aqueous redispersion of core–shell type nanocomposite particles produced via fluidized bed coating. *Powder Technology,* 236**,** 221-234.

BREDA, S. A., JIMENEZ-KAIRUZ, A. F., MANZO, R. H. & OLIVERA, M. E. 2009. Solubility behavior and biopharmaceutical classification of novel high-solubility ciprofloxacin and norfloxacin pharmaceutical derivatives. *International Journal of Pharmaceutics,* 371**,** 106-13.

CALLEGARI, E., MALHOTRA, B., BUNGAY, P. J., WEBSTER, R., FENNER, K. S., KEMPSHALL, S., LAPERLE, J. L., MICHEL, M. C. & KAY, G. G. 2011. A comprehensive non-clinical evaluation of the CNS penetration potential of antimuscarinic agents for the treatment of overactive bladder. *British Journal of Clinical Pharmacology,* 72**,** 235-46.

CAO, D., WANG, J., ZHOU, R., LI, Y., YU, H. & HOU, T. 2012. ADMET evaluation in drug discovery. 11. PharmacoKinetics Knowledge Base (PKKB): a comprehensive database of pharmacokinetic and toxic properties for drugs. *Journal of Chemical Information and Modeling,* 52**,** 1132-7.

CATALANO, A., DESAPHY, J. F., LENTINI, G., CAROCCI, A., DI MOLA, A., BRUNO, C., CARBONARA, R., DE PALMA, A., BUDRIESI, R., GHELARDINI, C., PERRONE, M. G., COLABUFO, N. A., CONTE CAMERINO, D. & FRANCHINI, C. 2012. Synthesis and toxicopharmacological evaluation of m-hydroxymexiletine,

the first metabolite of mexiletine more potent than the parent compound on voltage-gated sodium channels. *Journal of Medicinal Chemistry,* 55**,** 1418-22.

CHEN, L. L., YAO, J., YANG, J. B. & YANG, J. 2005. Predicting MDCK cell permeation coefficients of organic molecules using membrane-interaction QSAR analysis. *Acta Pharmacologica Sinica,* 26**,** 1322-33.

CHOI, M. K. & SONG, I. S. 2012. Characterization of efflux transport of the PDE5 inhibitors, vardenafil and sildenafil. *Journal of Pharmacy and Pharmacology,* 64**,** 1074-1083.

CHONG, S., DANDO, S. A., SOUCEK, K. M. & MORRISON, R. A. 1996. In vitro permeability through caco-2 cells is not quantitatively predictive of in vivo absorption for peptide-like drugs absorbed via the dipeptide transporter system. *Pharmaceutical Research,* 13**,** 120-3.

CLARYSSE, S., PSACHOULIAS, D., BROUWERS, J., TACK, J., ANNAERT, P., DUCHATEAU, G., REPPAS, C. & AUGUSTIJNS, P. 2009. Postprandial changes in solubilizing capacity of human intestinal fluids for BCS class II drugs. *Pharmaceutical Research,* 26**,** 1456-66.

COLLETT, A., SIMS, E., WALKER, D., HE, Y. L., AYRTON, J., ROWLAND, M. & WARHURST, G. 1996. Comparison of HT29-18-C1 and Caco-2 cell lines as models for studying intestinal paracellular drug absorption. *Pharmaceutical Research,* 13**,** 216-21.

COOK, J., ADDICKS, W. & WU, Y. H. 2008. Application of the biopharmaceutical classification system in clinical drug development--an industrial view. *The AAPS Journal,* 10**,** 306-10.

CRIVORI, P., REINACH, B., PEZZETTA, D. & POGGESI, I. 2006. Computational models for identifying potential P-glycoprotein substrates and inhibitors. *Molecular Pharmaceutics,* 3**,** 33-44.

CROWE, A. & TEOH, Y. K. 2006. Limited P-glycoprotein mediated efflux for anti-epileptic drugs. *Journal of Drug Targeting,* 14**,** 291-300.

CROWE, A. & WONG, P. 2003. Potential roles of P-gp and calcium channels in loperamide and diphenoxylate transport. *Toxicology and Applied Pharmacology,* 193**,** 127-37.

DAMLE, B., VARMA, M. V. & WOOD, N. 2011. Pharmacokinetics of voriconazole administered concomitantly with fluconazole and population-based simulation for sequential use. *Antimicrobial Agents and Chemotherapy,* 55**,** 5172-7.

DI, L., WHITNEY-PICKETT, C., UMLAND, J. P., ZHANG, H., ZHANG, X., GEBHARD, D. F., LAI, Y. R., FEDERICO, J. J., DAVIDSON, R. E., SMITH, R., REYNER, E. L., LEE, C., FENG, B., ROTTER, C., VARMA, M. V., KEMPSHALL, S., FENNER, K., EL-KATTAN, A. F., LISTON, T. E. & TROUTMAN, M. D. 2011. Development of a New Permeability Assay Using Low-Efflux MDCKII Cells. *Journal of Pharmaceutical Sciences,* 100**,** 4974-4985.

DIXIT, M., KINI, A. G. & KULKARNI, P. K. 2011. Enhancing the aqueous solubility and dissolution of olanzapine using freeze-drying. *Brazilian Journal of Pharmaceutical Sciences,* 47**,** 743-749.

DOW, J., FRANCESCO, G. F. & BERG, C. 1996. Comparison of the pharmacokinetics of dolasetron and its major active metabolite, reduced dolasetron, in dog. *Journal of Pharmaceutical Sciences,* 85**,** 685-9.

DRESSMAN, J. B., NAIR, A., ABRAHAMSSON, B., BARENDS, D. M., GROOT, D. W., KOPP, S., LANGGUTH, P., POLLI, J. E., SHAH, V. P. & ZIMMER, M. 2012. Biowaiver monograph for immediate-release solid oral dosage forms: acetylsalicylic acid. *Journal of Pharmaceutical Sciences,* 101**,** 2653-67.

DYCK, B., TAMIYA, J., JOVIC, F., PICK, R. R., BRADBURY, M. J., O'BRIEN, J., WEN, J., JOHNS, M., MADAN, A., FLECK, B. A., FOSTER, A. C., LI, B., ZHANG, M., TRAN, J. A., VICKERS, T., GREY, J., SAUNDERS, J. & CHEN, C. 2008. Characterization of thien-2-yl 1S,2R-milnacipran analogues as potent

norepinephrine/serotonin transporter inhibitors for the treatment of neuropathic pain. *Journal of Medicinal Chemistry,* 51**,** 7265-72.

ELSHAER, A., KHAN, S., PERUMAL, D., HANSON, P. & MOHAMMED, A. R. 2011. Use of amino acids as counterions improves the solubility of the BCS II model drug, indomethacin. *Current Drug Delivery,* 8**,** 363-72.

FAASSEN, F., KELDER, J., LENDERS, J., ONDERWATER, R. & VROMANS, H. 2003. Physicochemical Properties and Transport of Steroids Across Caco-2 Cells. *Pharmaceutical Research,* 20**,** 177-186.

FENG, B., MILLS, J. B., DAVIDSON, R. E., MIRELES, R. J., JANISZEWSKI, J. S., TROUTMAN, M. D. & DE MORAIS, S. M. 2008. In vitro P-glycoprotein assays to predict the in vivo interactions of P-glycoprotein with drugs in the central nervous system. *Drug Metabolism and Disposition,* 36**,** 268-75.

FRICK, A., MOLLER, H. & WIRBITZKI, E. 1998. Biopharmaceutical characterization of oral immediate release drug products. In vitro/in vivo comparison of phenoxymethylpenicillin potassium, glimepiride and levofloxacin. *European Journal of Pharmaceutics and Biopharmaceutics,* 46**,** 305-11.

GARBERG, P., BALL, M., BORG, N., CECCHELLI, R., FENART, L., HURST, R. D., LINDMARK, T., MABONDZO, A., NILSSON, J. E., RAUB, T. J., STANIMIROVIC, D., TERASAKI, T., ÖBERG, J. O. & ÖSTERBERG, T. 2005. In vitro models for the blood–brain barrier. *Toxicology In Vitro,* 19**,** 299-334.

GNOTH, M. J., BUETEHORN, U., MUENSTER, U., SCHWARZ, T. & SANDMANN, S. 2011. In vitro and in vivo P-glycoprotein transport characteristics of rivaroxaban. *Journal of Pharmacology and Experimental Therapeutics,* 338**,** 372-80.

GONZALEZ-ESQUIVEL, D., RIVERA, J., CASTRO, N., YEPEZ-MULIA, L. & JUNG COOK, H. 2005. In vitro characterization of some biopharmaceutical properties of praziquantel. *International Journal of Pharmaceutics,* 295**,** 93-9.

GRANERO, G. E., LONGHI, M. R., BECKER, C., JUNGINGER, H. E., KOPP, S., MIDHA, K. K., SHAH, V. P., STAVCHANSKY, S., DRESSMAN, J. B. & BARENDS, D. M. 2008. Biowaiver monographs for immediate release solid oral dosage forms: acetazolamide. *Journal of Pharmaceutical Sciences,* 97**,** 3691-9.

GRANERO, G. E., LONGHI, M. R., MORA, M. J., JUNGINGER, H. E., MIDHA, K. K., SHAH, V. P., STAVCHANSKY, S., DRESSMAN, J. B. & BARENDS, D. M. 2010. Biowaiver Monographs for Immediate Release Solid Oral Dosage Forms: Furosemide. *Journal of Pharmaceutical Sciences,* 99**,** 2544-2556.

GRBIC, S., PAROJCIC, J., IBRIC, S. & DJURIC, Z. 2011. In vitro-in vivo correlation for gliclazide immediate-release tablets based on mechanistic absorption simulation. *AAPS PharmSciTech,* 12**,** 165-71.

GRES, M. C., JULIAN, B., BOURRIE, M., MEUNIER, V., ROQUES, C., BERGER, M., BOULENC, X., BERGER, Y. & FABRE, G. 1998. Correlation between oral drug absorption in humans, and apparent drug permeability in TC-7 cells, a human epithelial intestinal cell line: Comparison with the parental Caco-2 cell line. *Pharmaceutical Research,* 15**,** 726-733.

HARRISON, A., BETTS, A., FENNER, K., BEAUMONT, K., EDGINGTON, A., ROFFEY, S., DAVIS, J., COMBY, P. & MORGAN, P. 2004. Non- linear oral pharmacokinetics of the alpha-antagonist 4-amino-5-(4-fluorophenyl)-6,7-dimethoxy-2-[4-(morpholinocarbonyl)-perhydro-1,4-d iazepin-1-yl]quinoline in humans: use of preclinical data to rationalize clinical observations. *Drug Metabolism and Disposition,* 32**,** 197-204.

HASSAN, H. E., MYERS, A. L., LEE, I. J., COOP, A. & EDDINGTON, N. D. 2007. Oxycodone induces overexpression of P-glycoprotein (ABCB1) and affects paclitaxel's tissue distribution in Sprague Dawley rats. *Journal of Pharmaceutical Sciences,* 96**,** 2494-506.

HE, H., TRAN, P., YIN, H., SMITH, H., FLOOD, D., KRAMP, R., FILIPECK, R., FISCHER, V. & HOWARD, D. 2009. Disposition of vildagliptin, a novel dipeptidyl peptidase 4 inhibitor, in rats and dogs. *Drug Metabolism and Disposition,* 37**,** 545-54.

HEMMINGSEN, P. H., HAAHR, A.-M., GUNNERGAARD, C. & CARDOT, J.-M. 2011. Development of a New Type of Prolonged Release Hydrocodone Formulation Based on Egalet® ADPREM Technology Using In Vivo–In Vitro Correlation. *Pharmaceutics,* 3**,** 73-87.

HILGENDORF, C., SPAHN-LANGGUTH, H., REGARDH, C. G., LIPKA, E., AMIDON, G. L. & LANGGUTH, P. 2000. Caco-2 versus Caco-2/HT29-MTX co-cultured cell lines: Permeabilities via diffusion, inside- and outside-directed carrier-mediated transport. *Journal of Pharmaceutical Sciences,* 89**,** 63-75.

HUANG, L., BERRY, L., GANGA, S., JANOSKY, B., CHEN, A., ROBERTS, J., COLLETTI, A. E. & LIN, M. H. 2010. Relationship between passive permeability, efflux, and predictability of clearance from in vitro metabolic intrinsic clearance. *Drug Metabolism and Disposition,* 38**,** 223-31.

INGELS, F., OTH, M., AUGUSTIJNS, P. 2004. Evaluation of the in vivo dissolution profile of orally administered theophylline in man (Abstract). *In:* Exposition, A. A. M. A., ed., November 7-11 2004 Baltimore.

IRVINE, J. D., TAKAHASHI, L., LOCKHART, K., CHEONG, J., TOLAN, J. W., SELICK, H. E. & GROVE, J. R. 1999. MDCK (Madin-Darby canine kidney) cells: A tool for membrane permeability screening. *Journal of Pharmaceutical Sciences,* 88**,** 28-33.

JANTRATID, E., PRAKONGPAN, S., DRESSMAN, J. B., AMIDON, G. L., JUNGINGER, H. E., MIDHA, K. K. & BARENDS, D. M. 2006. Biowaiver monographs for immediate release solid oral dosage forms: cimetidine. *Journal of Pharmaceutical Sciences,* 95**,** 974-84.

JEZYK, N., LI, C., STEWART, B. H., WU, X., BOCKBRADER, H. N. & FLEISHER, D. 1999. Transport of pregabalin in rat intestine and Caco-2 monolayers. *Pharmaceutical Research,* 16**,** 519-26.

JIN, H. E., SONG, B., KIM, S. B., SHIM, W. S., KIM, D. D., CHONG, S., CHUNG, S. J. & SHIM, C. K. 2013. Transport of gemifloxacin, a 4th generation quinolone antibiotic, in the Caco-2 and engineered MDCKII cells, and potential involvement of efflux transporters in the intestinal absorption of the drug. *Xenobiotica,* 43**,** 355-67.

JUNG, S. J., CHOI, S. O., UM, S. Y., KIM, J. I., CHOO, H. Y. P., CHOI, S. Y. & CHUNG, S. Y. 2006. Prediction of the permeability of drugs through study on quantitative structure–permeability relationship. *Journal of Pharmaceutical and Biomedical Analysis,* 41**,** 469-475.

KALANTZI, L., REPPAS, C., DRESSMAN, J., AMIDON, G., JUNGINGER, H., MIDHA, K., SHAH, V., STAVCHANSKY, S. & BARENDS, D. M. 2006. Biowaiver monographs for immediate release solid oral dosage forms: Acetaminophen (paracetamol). *Journal of Pharmaceutical Sciences,* 95**,** 4-14.

KANAAN, M., DAALI, Y., DAYER, P. & DESMEULES, J. 2009. P-glycoprotein is not involved in the differential oral potency of naloxone and naltrexone. *Fundam Clin Pharmacol,* 23**,** 543-8.

KATDARE, A. & CHAUBAL, M. 2006. *Excipient Development for Pharmaceutical, Biotechnology, and Drug Delivery Systems,* New York, CRC Press.

KHAN, M. Z., RAUSL, D., ZANOSKI, R., ZIDAR, S., MIKULCIC, J. H., KRIZMANIC, L., ESKINJA, M., MILDNER, B. & KNEZEVIC, Z. 2004. Classification of loratadine based on the biopharmaceutics drug classification concept and possible in vitro-in vivo correlation. *Biological and Pharmaceutical Bulletin,* 27**,** 1630-5.

KHAN, S., BATCHELOR, H., HANSON, P., PERRIE, Y. & MOHAMMED, A. R. 2011. Physicochemical characterisation, drug polymer dissolution and in vitro evaluation

of phenacetin and phenylbutazone solid dispersions with polyethylene glycol 8000. *Journal of Pharmaceutical Sciences,* 10**,** 22613.

KI, M. H., CHOI, M. H., AHN, K. B., KIM, B. S., IM, D. S., AHN, S. K. & SHIN, H. J. 2008. The efficacy and safety of clopidogrel resinate as a novel polymeric salt form of clopidogrel. *Archives of Pharmaceutical Research,* 31**,** 250-8.

KIKUCHI, R., DE MORAIS, S. M. & KALVASS, J. C. 2013. In vitro P-glycoprotein efflux ratio can predict the in vivo brain penetration regardless of biopharmaceutics drug disposition classification system class. *Drug Metabolism and Disposition,* 41**,** 2012-7.

KIM, W. Y. & BENET, L. Z. 2004. P-glycoprotein (P-gp/MDR1)-mediated efflux of sex-steroid hormones and modulation of P-gp expression in vitro. *Pharmaceutical Research,* 21**,** 1284-93.

KIS, O., ZASTRE, J. A., HOQUE, M. T., WALMSLEY, S. L. & BENDAYAN, R. 2013. Role of drug efflux and uptake transporters in atazanavir intestinal permeability and drug-drug interactions. *Pharmaceutical Research,* 30**,** 1050-64.

KOGAN, A., KESSELMAN, E., DANINO, D., ASERIN, A. & GARTI, N. 2008. Viability and permeability across Caco-2 cells of CBZ solubilized in fully dilutable microemulsions. *Colloids and Surfaces B: Biointerfaces,* 66**,** 1-12.

KOVACEVIC, I., PAROJCIC, J., HOMSEK, I., TUBIC-GROZDANIS, M. & LANGGUTH, P. 2009. Justification of biowaiver for carbamazepine, a low soluble high permeable compound, in solid dosage forms based on IVIVC and gastrointestinal simulation. *Molecular Pharmaceutics,* 6**,** 40-7.

KUO, K.-L., ZHU, H., MCNAMARA, P. J. & LEGGAS, M. 2012. Localization and Functional Characterization of the Rat Oatp4c1 Transporter in an In Vitro Cell System and Rat Tissues. *PLOS One,* 7.

LAITINEN, L., KANGAS, H., KAUKONEN, A. M., HAKALA, K., KOTIAHO, T., KOSTIAINEN, R. & HIRVONEN, J. 2003. N-in-one permeability studies of heterogeneous sets of compounds across Caco-2 cell monolayers. *Pharmaceutical Research,* 20**,** 187-97.

LANGGUTH, P., KUBIS, A., KRUMBIEGEL, G., LANG, W., MERKLE, H. P., WÄCHTER, W., SPAHN-LANGGUTH, H. & WEYHENMEYER, R. 1997. Intestinal absorption of the quaternary trospium chloride: permeability-lowering factors and bioavailabilities for oral dosage forms. *European Journal of Pharmaceutics and Biopharmaceutics,* 43**,** 265-272.

LASSOUED, M. A., KHEMISS, F. & SFAR, S. 2011. Comparative study of two in vitro methods for assessing drug absorption: Sartorius SM 16750 apparatus versus Everted Gut Sac. *Journal of Pharmaceutical Sciences,* 14**,** 117-27.

LEVITT, D. G. 2013. Quantitation of small intestinal permeability during normal human drug absorption. *BMC Pharmacology and Toxicology,* 14**,** 2050-6511.

LI, F., HONG, L., MAU, C. I., CHAN, R., HENDRICKS, T., DVORAK, C., YEE, C., HARRIS, J. & ALFREDSON, T. 2006. Transport of levovirin prodrugs in the human intestinal Caco-2 cell line. *Journal of Pharmaceutical Sciences,* 95**,** 1318-25.

LINDENBERG, M., KOPP, S. & DRESSMAN, J. B. 2004. Classification of orally administered drugs on the World Health Organization Model list of Essential Medicines according to the biopharmaceutics classification system. *European Journal of Pharmaceutics and Biopharmaceutics,* 58**,** 265-278.

LOSCHER, W. & POTSCHKA, H. 2005. Blood-brain barrier active efflux transporters: ATP-binding cassette gene family. *NeuroRx,* 2**,** 86-98.

LU, Z., CHEN, W., VILJOEN, A. & HAMMAN, J. H. 2010. Effect of sinomenine on the in vitro intestinal epithelial transport of selected compounds. *Phytotherapy Research,* 24**,** 211-8.

LUO, S., WANG, Z., KANSARA, V., PAL, D. & MITRA, A. K. 2008. Activity of a sodium-dependent vitamin C transporter (SVCT) in MDCK-MDR1 cells and

mechanism of ascorbate uptake. *International Journal of Pharmaceutics,* 358**,** 168-76.

MAHAR DOAN, K. M., HUMPHREYS, J. E., WEBSTER, L. O., WRING, S. A., SHAMPINE, L. J., SERABJIT-SINGH, C. J., ADKISON, K. K. & POLLI, J. W. 2002. Passive permeability and P-glycoprotein-mediated efflux differentiate central nervous system (CNS) and non-CNS marketed drugs. *Journal of Pharmacology and Experimental Therapeutics,* 303**,** 1029-37.

MAINES, L. W., ANTONETTI, D. A., WOLPERT, E. B. & SMITH, C. D. 2005. Evaluation of the role of P-glycoprotein in the uptake of paroxetine, clozapine, phenytoin and carbamazapine by bovine retinal endothelial cells. *Neuropharmacology,* 49**,** 610-7.

MARASANAPALLE, V. P., CRISON, J. R., MA, J., LI, X. & JASTI, B. R. 2009. Investigation of some factors contributing to negative food effects. *Biopharmaceutics & Drug Disposition,* 30**,** 71-80.

MARTINDALE 2009. *Martindale The Complete Drug Reference,* London, Pharmaceutical Press.

MATHIAS, A., MENNING, M., WISER, L., WEI, X. & DAVE, A. 2012. Bioequivalence of the Emtricitabine/Rilpivirine/Tenofovir Disoproxil Fumarate Single Tablet Regimen. *Journal of Bioequivalence & Bioavailability,* 4**,** 100-105.

MILNER, E., MCCALMONT, W., BHONSLE, J., CARIDHA, D., COBAR, J., GARDNER, S., GERENA, L., GOODINE, D., LANTERI, C., MELENDEZ, V., RONCAL, N., SOUSA, J., WIPF, P. & DOW, G. S. 2010. Anti-malarial activity of a non-piperidine library of next-generation quinoline methanols. *Malaria J,* 9**,** 1475-2875.

MONTEIRO, L. M., LIONE, V. F., DO CARMO, F. A., DO AMARAL, L. H., DA SILVA, J. H., NASCIUTTI, L. E., RODRIGUES, C. R., CASTRO, H. C., DE SOUSA, V. P. & CABRAL, L. M. 2012. Development and characterization of a new oral dapsone nanoemulsion system: permeability and in silico bioavailability studies. *International Journal of Nanomedicine,* 7**,** 5175-82.

NAGPAL, M., RAJERA, R., NAGPAL, K., RAKHA, P., SINGH, S. & MISHRA, D. 2012. Dissolution enhancement of glimepiride using modified gum karaya as a carrier. *International Journal of Pharmaceutical Investigation,* 2**,** 42-7.

NAINAR, S., RAJIAH, K., ANGAMUTHU, S., PRABAKARAN, D. & KASIBHATTA, R. 2012. Biopharmaceutical Classification System in Invitro/ In-vivo Correlation: Concept and Development Strategies in Drug Delivery. *Tropical Journal of Pharmaceutical Research,* 11**,** 319-329.

NIELSEN, C. U., FROLUND, S., ABDULHADI, S., SARI, H., LANGTHALER, L., NOHR, M. K., KALL, M. A., BRODIN, B. & HOLM, R. 2013. Sertraline inhibits the transport of PAT1 substrates in vivo and in vitro. *British Journal of Pharmacology,* 170**,** 1041-52.

NORDQVIST, A., NILSSON, J., LINDMARK, T., ERIKSSON, A., GARBERG, P. & KIHLÉN, M. 2004. A General Model for Prediction of Caco-2 Cell Permeability. *QSAR & Combinatorial Science,* 23**,** 303-310.

OO, C., SNELL, P., BARRETT, J., DORR, A., LIU, B. & WILDING, I. 2003. Pharmacokinetics and delivery of the anti-influenza prodrug oseltamivir to the small intestine and colon using site-specific delivery capsules. *International Journal of Pharmaceutics,* 257**,** 297-9.

PALAPARTHY, R., PRADHAN, R. S., CHAN, J., RIESER, M., CHIRA, T., GALITZ, L., AWNI, W. & WILLIAMS, L. A. 2007. Effect of omeprazole on the pharmacokinetics of paricalcitol in healthy subjects. *Biopharmaceutics & Drug Disposition,* 28**,** 65-71.

PAPAGEORGIOU, G., DOCOSLIS, A., GEORGARAKIS, M. & BIKIARIS, D. 2009. The effect of physical state on the drug dissolution rate. *Journal of Thermal Analysis and Calorimetry,* 95**,** 903-915.

POLLI, J. W., WRING, S. A., HUMPHREYS, J. E., HUANG, L., MORGAN, J. B., WEBSTER, L. O. & SERABJIT-SINGH, C. S. 2001. Rational use of in vitro P-glycoprotein assays in drug discovery. *Journal of Pharmacology and Experimental Therapeutics,* 299**,** 620-8.

RADWAN, A., AMIDON, G. L. & LANGGUTH, P. 2012. Mechanistic investigation of food effect on disintegration and dissolution of BCS class III compound solid formulations: the importance of viscosity. *Biopharmaceutics & Drug Disposition,* 33**,** 403-16.

RAEISSI, S. D., LI, J. & HIDALGO, I. J. 1999. The role of an alpha-amino group on H+ -dependent transepithelial transport of cephalosporins in Caco-2 cells. *Journal of Pharmacy and Pharmacology,* 51**,** 35-40.

RAMIREZ, E., LAOSA, O., GUERRA, P., DUQUE, B., MOSQUERA, B., BOROBIA, A. M., LEI, S. H., CARCAS, A. J. & FRIAS, J. 2010. Acceptability and characteristics of 124 human bioequivalence studies with active substances classified according to the Biopharmaceutic Classification System. *British Journal of Clinical Pharmacology,* 70**,** 694-702.

RANALDI, G., ISLAM, K. & SAMBUY, Y. 1992. Epithelial cells in culture as a model for the intestinal transport of antimicrobial agents. *Antimicrobial Agents and Chemotherapy,* 36**,** 1374-81.

REGE, B. D., YU, L. X., HUSSAIN, A. S. & POLLI, J. E. 2001. Effect of common excipients on Caco-2 transport of low-permeability drugs. *Journal of Pharmaceutical Sciences,* 90**,** 1776-86.

SHAIK, N., GIRI, N., PAN, G. & ELMQUIST, W. F. 2007. P-glycoprotein-mediated active efflux of the anti-HIV1 nucleoside abacavir limits cellular accumulation and brain distribution. *Drug Metabolism and Disposition,* 35**,** 2076-85.

SHARMA, P., VARMA, M. V., CHAWLA, H. P. & PANCHAGNULA, R. 2005. Relationship between lipophilicity of BCS class III and IV drugs and the functional activity of peroral absorption enhancers. *Il Farmaco,* 60**,** 870-3.

SHERER, E. C., VERRAS, A., MADEIRA, M., HAGMANN, W. K., SHERIDAN, R. P., ROBERTS, D., BLEASBY, K. & CORNELL, W. D. 2012. QSAR Prediction of Passive Permeability in the LLC-PK1 Cell Line: Trends in Molecular Properties and Cross-Prediction of Caco-2 Permeabilities. *Molecular Informatics,* 31**,** 231-245.

SINGLETON, D. H., BOYD, H., STEIDL-NICHOLS, J. V., DEACON, M., GROOT, M. J., PRICE, D., NETTLETON, D. O., WALLACE, N. K., TROUTMAN, M. D., WILLIAMS, C. & BOYD, J. G. 2007. Fluorescently labeled analogues of dofetilide as high-affinity fluorescence polarization ligands for the human ether-a-go-go-related gene (hERG) channel. *Journal of Medicnal Chemistry,* 50**,** 2931-41.

SKERJANEC, A. 2006. The clinical pharmacokinetics of darifenacin. *Clinical Pharmacokinetics,* 45**,** 325-50.

SKOLD, C., WINIWARTER, S., WERNEVIK, J., BERGSTROM, F., ENGSTROM, L., ALLEN, R., BOX, K., COMER, J., MOLE, J., HALLBERG, A., LENNERNAS, H., LUNDSTEDT, T., UNGELL, A. L. & KARLEN, A. 2006. Presentation of a structurally diverse and commercially available drug dataset for correlation and benchmarking studies. *Journal of Medicnal Chemistry,* 49**,** 6660-71.

SKOLNIK, S., LIN, X., WANG, J., CHEN, X. H., HE, T. & ZHANG, B. 2010. Towards prediction of in vivo intestinal absorption using a 96-well Caco-2 assay. *Journal of Pharmaceutical Sciences,* 99**,** 3246-65.

SMETANOVA, L., STETINOVA, V., KHOLOVA, D., KVETINA, J., SMETANA, J. & SVOBODA, Z. 2009. Caco-2 cells and Biopharmaceutics Classification System

(BCS) for prediction of transepithelial transport of xenobiotics (model drug: caffeine). *Neuroendocrinology Letters,* 1**,** 101-5.

STENBERG, P., NORINDER, U., LUTHMAN, K. & ARTURSSON, P. 2001. Experimental and computational screening models for the prediction of intestinal drug absorption. *Journal of Medicinal Chemistry,* 44**,** 1927-1937.

SU, T. Z., FENG, M. R. & WEBER, M. L. 2005. Mediation of highly concentrative uptake of pregabalin by L-type amino acid transport in Chinese hamster ovary and Caco-2 cells. *Journal of Pharmacology and Experimental Therapeutics,* 313**,** 1406-15.

SUBRAMANIAN, S., SINGIREDDY, A., KRISHNAMOORTHY, K. & RAJAPPAN, M. 2012. Nanosponges: a novel class of drug delivery system--review. *Journal of Pharmaceutical Sciences,* 15**,** 103-11.

TAUPITZ, T., DRESSMAN, J. B. & KLEIN, S. 2013. New formulation approaches to improve solubility and drug release from fixed dose combinations: case examples pioglitazone/glimepiride and ezetimibe/simvastatin. *European Journal of Pharmaceutics and Biopharmaceutics,* 84**,** 208-18.

THAKURIA, R. & NANGIA, A. 2011. Highly soluble olanzapinium maleate crystalline salts. *CrystEngComm,* 13**,** 1759-1764.

TSUJI, A., TERASAKI, T., TAMAI, I. & HIROOKA, H. 1987. H+ gradient-dependent and carrier-mediated transport of cefixime, a new cephalosporin antibiotic, across brush-border membrane vesicles from rat small intestine. *Journal of Pharmacology and Experimental Therapeutics,* 241**,** 594-601.

TSUME, Y. & AMIDON, G. L. 2010. The biowaiver extension for BCS class III drugs: the effect of dissolution rate on the bioequivalence of BCS class III immediate-release drugs predicted by computer simulation. *Molecular Pharmaceutics,* 7**,** 1235-43.

VARMA, M. V., GARDNER, I., STEYN, S. J., NKANSAH, P., ROTTER, C. J., WHITNEY-PICKETT, C., ZHANG, H., DI, L., CRAM, M., FENNER, K. S. & EL-KATTAN, A. F. 2012. pH-Dependent Solubility and Permeability Criteria for Provisional Biopharmaceutics Classification (BCS and BDDCS) in Early Drug Discovery. *Molecular Pharmaceutics,* 9**,** 1199-1212.

VARMA, M. V. S., SATEESH, K. & PANCHAGNULA, R. 2005. Functional role of P-glycoprotein in limiting intestinal absorption of drugs: Contribution of passive permeability to P-glycoprotein mediated efflux transport. *Molecular Pharmaceutics,* 2**,** 12-21.

VIOLETTE, A., CORTES, D. A., BERGEON, J. A., FALCONER, R. A. & TOTH, I. 2008. Optimized LC-MS/MS quantification method for the detection of piperacillin and application to the development of charged liposaccharides as oral penetration enhancers. *International Journal of Pharmaceutics,* 351**,** 152-7.

WAGER, T. T., CHANDRASEKARAN, R. Y., HOU, X., TROUTMAN, M. D., VERHOEST, P. R., VILLALOBOS, A. & WILL, Y. 2010. Defining Desirable Central Nervous System Drug Space through the Alignment of Molecular Properties, in Vitro ADME, and Safety Attributes. *ACS Chemical Neuroscience,* 1**,** 420-434.

WANG, J. S., ZHU, H. J., GIBSON, B. B., MARKOWITZ, J. S., DONOVAN, J. L. & DEVANE, C. L. 2008. Sertraline and its metabolite desmethylsertraline, but not bupropion or its three major metabolites, have high affinity for P-glycoprotein. *Biological and Pharmaceutical Bulletin,* 31**,** 231-4.

WEI, H. & LOBENBERG, R. 2006. Biorelevant dissolution media as a predictive tool for glyburide a class II drug. *Eur Journal of Pharmaceutical Sciences,* 29**,** 45-52.

WENZEL, U., KUNTZ, S., DIESTEL, S. & DANIEL, H. 2002. PEPT1-mediated cefixime uptake into human intestinal epithelial cells is increased by Ca2+ channel blockers. *Antimicrobial Agents and Chemotherapy,* 46**,** 1375-80.

WHO 2006. Annex 8: Proposal to waive in vivo bioequivalence requirements for WHO Model List of Essential Medicines immediate-release, solid oral dosage forms;

Technical Report Series No. 937; 40th; WHO Expert Committee on Specification for Pharmaceutical Preparations, WHO Technical Report Series, No. 937, 2006; pp 391–461
http://www.who.int/medicines/publications/essentialmedicines/en/index.html (accessed December 19, 2013).

WU, C. Y. & BENET, L. Z. 2005. Predicting drug disposition via application of BCS: Transport/absorption/elimination interplay and development of a biopharmaceutics drug disposition classification system. *Pharmaceutical Research,* 22**,** 11-23.

YAMASHITA, S., FURUBAYASHI, T., KATAOKA, M., SAKANE, T., SEZAKI, H. & TOKUDA, H. 2000. Optimized conditions for prediction of intestinal drug permeability using Caco-2 cells. *European Journal of Pharmaceutical Sciences,* 10**,** 195-204.

YANEZ, J. A., REMSBERG, C. M., SAYRE, C. L., FORREST, M. L. & DAVIES, N. M. 2011. Flip-flop pharmacokinetics--delivering a reversal of disposition: challenges and opportunities during drug development. *Therapeutic delivery,* 2**,** 643-72.

YAZDANIAN, M., GLYNN, S. L., WRIGHT, J. L. & HAWI, A. 1998. Correlating partitioning and caco-2 cell permeability of structurally diverse small molecular weight compounds. *Pharmaceutical Research,* 15, 1490-4.

YEE, S. Y. 1997. In vitro permeability across Caco3 cells (colonic) can predict in vivo (small intestinal) absorption in man - Fact or myth. *Pharmaceutical Research,* 14**,** 763-766.

YOUNG, A. M., AUDUS, K. L., PROUDFOOT, J. & YAZDANIAN, M. 2006. Tetrazole compounds: the effect of structure and pH on Caco-2 cell permeability. *Journal of Pharmaceutical Sciences,* 95**,** 717-25.

ZERROUK, N., CORTI, G., ANCILLOTTI, S., MAESTRELLI, F., CIRRI, M. & MURA, P. 2006. Influence of cyclodextrins and chitosan, separately or in combination, on glyburide solubility and permeability. *European Journal of Pharmaceutics and Biopharmaceutics,* 62**,** 241-6.

ZORNOZA, T., CANO-CEBRIAN, M. J., NALDA-MOLINA, R., GUERRI, C., GRANERO, L. & POLACHE, A. 2004. Assessment and modulation of acamprosate intestinal absorption: comparative studies using in situ, in vitro (CACO-2 cell monolayers) and in vivo models. *European Journal of Pharmaceutical Sciences,* 22, 347-56.

Figures A4.1 – 3 Scatterplots between the first and second principal components of PCA for the solubility and permeability datasets

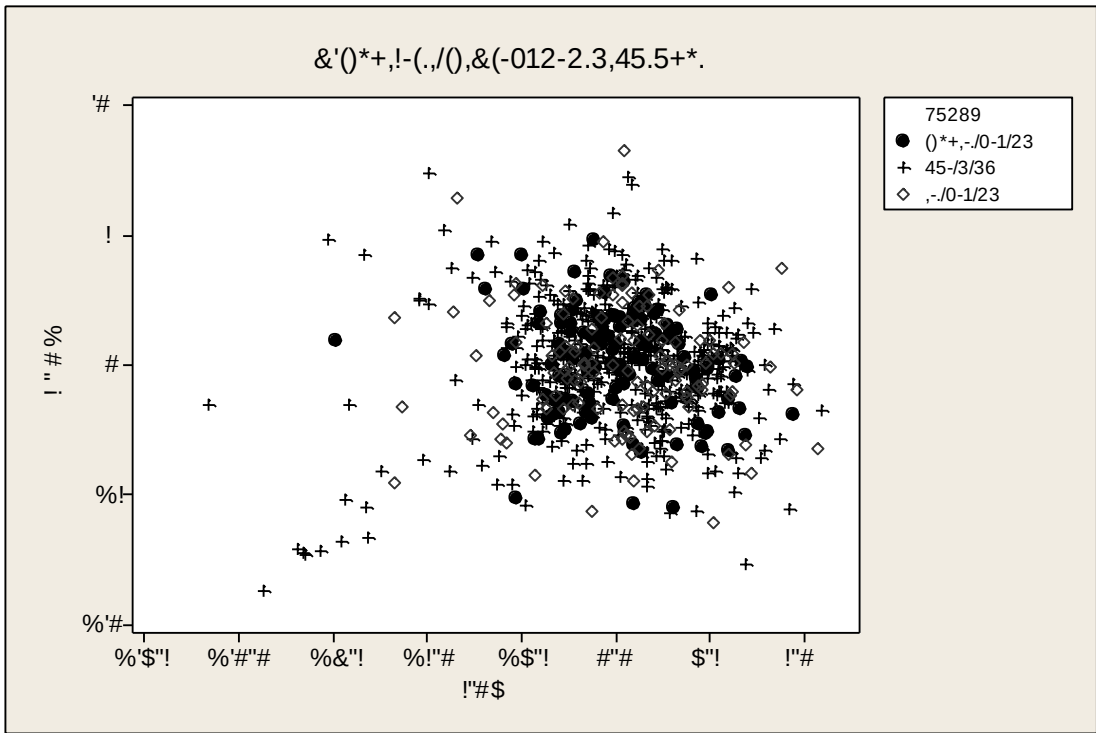

Figure A4. 1. The scatterplot between the first and second principal components of PCA for the solubility dataset

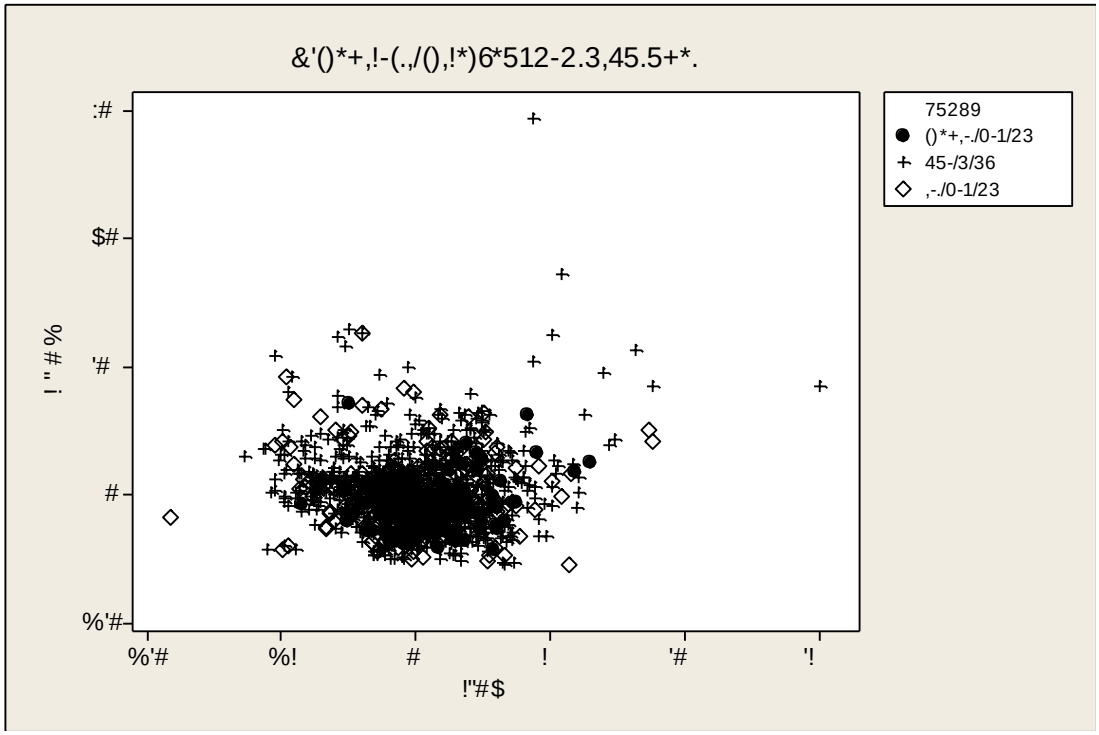

Figure A4. 2. The scatterplot between the first and second principal components of PCA for the permeability dataset

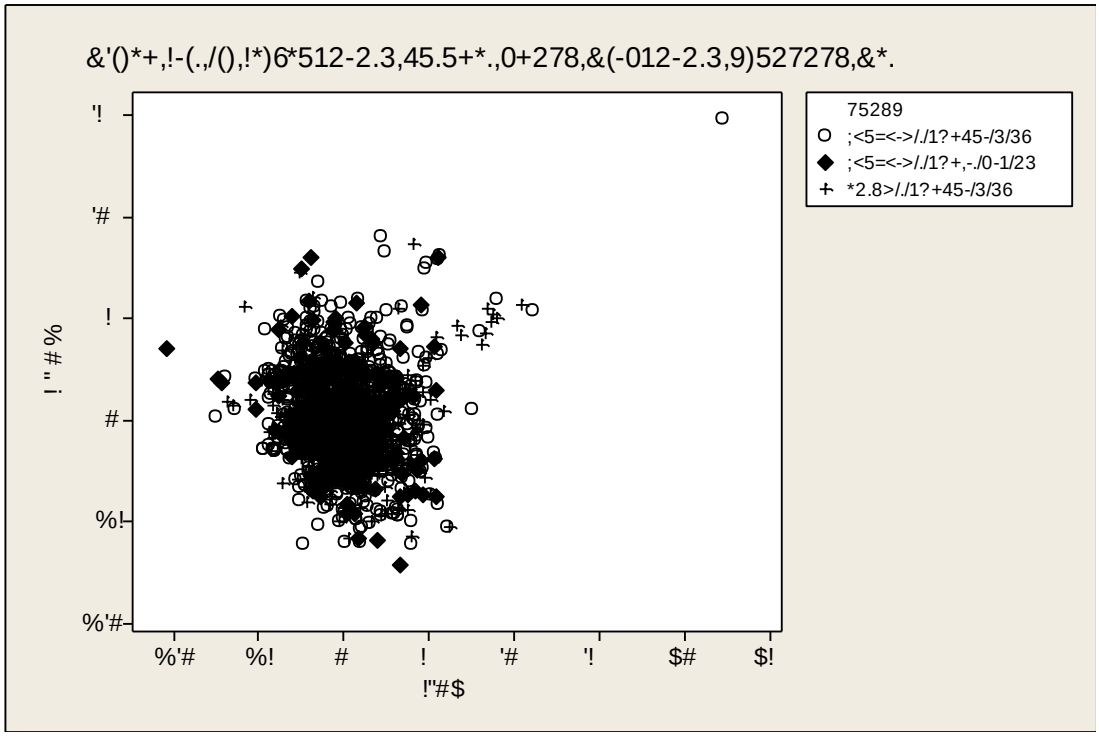

Figure A4. 3. The scatterplot between the first and second principal components of PCA for the permeability dataset using the solubility training set and top 20 molecular descriptors

Table A4. 4. List of 20 compounds plus SMILES in the permeability dataset determined to be outside the applicability domain for the solubility training set

Compound Name	Permeability dataset	SMILES
PAMAM-NH2 (G1)	t	O=C(NCCN(CCC(NCCN)=O)CCC(NCCN)=O)CCN(CCC(NCCN(CCC(NCCN)=O)CCC(NCCN)=O)=O)CCN(CCC(NCCN(CCC(NCCN)=O)CCC(NCCN)=O)=O)CCC(NCCN(CCC(NCCN)=O)CCC(NCCN)=O)=O
Ardeparin	t	CC(NC1C(O)C(OC2C(OS(=O)(O)=O)C(O)C(OC3C(NS(=O)(O)=O)C(OS(=O)(O)=O)C(OC4C(OS(=O)(O)=O)C(O)C(O)C(C(O)=O)O4)C(CO)O3)C(C(O)=O)O2)C(COS(=O)(O)=O)OC1O)=O
Phenazopyridine	t	CCC(C1C(NC(CCC(N)=O)C(NC(CC(N)=O)C(NC(C(N2CCCC2C(NC(C(NCC(N)=O)=O)CCC/N=C(N)\N)=O)=O)CSSCC(N)C(NC(CC3=CC=C(O)C=C3)C(N1)=O)=O)=O)=O)=O)C
Vasopressin	V	CCC(C1C(NC(CCC(N)=O)C(NC(CC(N)=O)C(NC(C(N2CCCC2C(NC(C(NCC(N)=O)=O)CCC/N=C(N)\N)=O)=O)CSSCC(N)C(NC(CC3=CC=C(O)C=C3)C(N1)=O)=O)=O)=O)=O)C
Arginine-vasopressin (AVP)	t	O=[C@@]([C@@H]1CSSC[C@H](N)C(N[C@H](CC2=CC=C(O)C=C2)C(N[C@H](CC3=CC=CC=C3)C(N[C@H](CCC(N)=O)C(N[C@@H](CC(N)=O)C(N1)=O)=O)=O)=O)=O)N4[C@@H]([C@@](N[C@H](C(NCC(N)=O)=O)CCC/N=C(N)/N)=O)CCC4
Deamino Arginine Vasopressin (dDAVP)	t	O=[C@@]([C@@H]1CSSCCC(N[C@@H](CC2=CC=C(O)C=C2)C(N[C@H](CC3=CC=CC=C3)C(N[C@@H](CCC(N)=O)C(N[C@@H](CC(N)=O)C(N1)=O)=O)=O)=O)=O)N4[C@H]([C@@](N[C@@H](C(NCC(N)=O)=O)CCC/N=C(N)/N)=O)CCC4
Desmopressin	t	O=C(C1CSSCCC(NC(CC2=CC=C(O)C=C2)C(NC(CC3=CC=CC=C3)C(NC(CCC(N)=O)C(NC(CC(N)=O)C(N1)=O)=O)=O)=O)=O)N4C(C(NC(C(NCC(N)=O)=O)CCC/N=C(N)\N)=O)CCC4
DALDA	t	CC1=C(C[C@H](N)C(N[C@@H](CCCNC(N)=N)C(NC(CC2=CC=CC=C2)C(NC(CCCCN)C(N)=O)=O)=O)=O)C(C)=CC(O)=C1
Human Beta-casomorphin-7 (Tyr-Pro-Phe-Val-Glu-Pro-Ile)	t	O=C(NC(C(N[C@@H](C(N1C(C(N[C@H]([C@@H](C)CC)C(O)=O)=O)CCC1)=O)CCC(O)=O)=O)C(C)C)[C@@H](NC(C2N(C([C@H](N)CC3=CC=C(O)C=C3)=O)CCC2)=O)CC4=CC=CC=C4
7	t	NC([C@@H](NC([C@H](CC1=CNC2=C1C=CC=C2)NC([C@@H]3N(C([C@H](CC4=CC=C(O)C=C4)NC(CCC(C[C@H]5[C@H]([C@H]([C@H](O)C(CO)O5)O)O)=O)=O)=O)=O)CCC3)=O)=O)CC6=CC=CC=C6)=O

Compound Name	Permeability dataset	SMILES
Rebaudioside A	t	C[C@@]12CCC[C@@](C)(C(OC3[C@@H](O)[C@H](O)[C@@H](O)[C@H](CO)O3)=O)[C@@]1([H])CC[C@@]45C2CC(O[C@@H]6[C@@H](O[C@@H]7[C@@H](O)[C@H](O)[C@@H](O)[C@H](CO)O7)C(O[C@@H]8[C@@H](O)[C@H](O)[C@@H](O)[C@H](CO)O8)[C@@H](O)[C@H](CO)O6)[C@H](C(C5)=C)C4
Leuprolide	t	CCNC(C1CCCN1C(C(NC(C(NC(C(NC(C(NC(C(NC(C(NC(C(N C(C2CCC(N2)=O)=O)CC3=CN=CN3)=O)CC4=CNC5=CC=CC=C54)=O)CO)=O)CC6=CC=C(O)C=C6)=O)CC(C)C)=O)CC(C)C)=O)CCC/N=C(N)\N)=O)=O
Buserelin	t	O=C1N[C@H]([C@](N[C@H](C(N[C@@H](CC2=CNC3=C2C=CC=C3)C(N[C@@H](CO)C(N[C@H](C(N[C@H](COC(C)(C)C)C(N[C@@H](CC(C)C)C(N[C@@H](CCC/N=C(N)/N)C(N4CCC[C@H]4[C@@](NCC)=O)=O)=O)=O)=O)CC5=CC=C(O)C=C5)=O)=O)=O)CC6=CN=CN6)=O)CC1
Amphotericin_B	BCS Val	C[C@H]1/C=C/C=C/C=C/C=C/C=C/C=C/C=C/C=C/[C@@H](C[C@H]2[C@@H]([C@H](C[C@](O2)(C[C@H](C[C@H]([C@@H](CC[C@H](C[C@H](CC(=O)O[C@H]([C@@H]([C@@H]1O)C)C)O)O)O)O)O)O)C(=O)O)O[C@H]3[C@H]([C@H]([C@@H]([C@H](O3)C)O)N)O
10	t	O=C(N[C@@H]1O[C@@H]([C@@H](O)C(O)C1O)C(N)=O)[C@@H](NC([C@H](CC2=CNC3=C2C=CC=C3)NC([C@@H]4N(C([C@H](CC5=CC=C(O)C=C5)N)=O)CCC4)=O)=O)CC6=CC=CC=C6
12	t	O=C(NC(C(N)=O)CCCCCC)[C@@H](NC([C@H](CC1=CNC2=C1C=CC=C2)NC([C@@H]3N(C([C@H](CC4=CC=C(O)C=C4)NC(CCC(N[C@H]5[C@H]([C@H]([C@H](O)C(CO)O5)O)O)=O)=O)=O)CCC3)=O)=O)CC6=CC=CC=C6
11	t	O=C(N[C@@H]1O[C@@H]([C@@H](O)C(O)C1O)C(N)=O)[C@@H](NC([C@H](CC2=CNC3=C2C=CC=C3)NC([C@@H]4N(C([C@H](CC5=CC=C(O)C=C5)NC(C(N)CCCCCC)=O)=O)CCC4)=O)=O)CC6=CC=CC=C6
Ginsenoside Rb1 (Rb1)	V	O[C@H]1[C@H](OC(CC/C=C(C)/C)([C@H]2CC[C@@]3(C)[C@]4(C)CC[C@@]5([H])C(C)(C)[C@@H](OC6O[C@H](CO)[C@@H](O)[C@H](O)[C@H]6O[C@H]7O[C@H](CO)[C@@H](O)[C@H](O)[C@H]7O)CC[C@]5(C)[C@@]4([H])C[C@@H](O)[C@]23[H])C)O[C@H](COC8O[C@H](CO)[C@@H](O)[C@H](O)[C@H]8O)[C@@H](O)[C@@H]1O
Actinomycin	t	CC1C(NC(C2=C3C(OC4=C(C)C(C(N)=C(C(NC5C(C)OC(C(C(C)C)N(C)C(CN(C)C([C@@H]6CCCN6C([C@@H](C(C)C)NC5=O)=O)=O)=O)=O)=O)C4=N3)=O)=C(C)C=C2)=O)C(N[C@H](C(C)C)C(N7CCC[C@H]7C(N(C)CC(N(C)C(C(C)C)C(O1)=O)=O)=O)=O)=O
14	t	NC([C@@H](NC([C@H](CC1=CNC2=C1C=CC=C2)NC([C@@H]3N(C([C@H](CC4=C(C)C=C(O)C=C4C)NC(C(N)CCCCCC)=O)=O)CCC3)=O)=O)CC5=CC=CC=C5)=O

T: training set, V: validation set, BCS Val: BCS external validation set

Publication List

The publications that have resulted from the research described in this thesis are detailed below. It comprises of published and submitted work for publication in scientific literature:

Publications

D. Newby, A. A Freitas and T. Ghafourian (2014b) Decision trees to characterise the roles of permeability and solubility on oral absorption. European Journal of Medicinal Chemistry (Accepted)

D. Newby, A. A Freitas and T. Ghafourian (2014a) Comparing Multi-label Classification Methods for Provisional Biopharmaceutics Class Prediction, Molecular Pharmaceutics (In Press)

D. Newby, A. A Freitas and T. Ghafourian (2013) Improving oral absorption models with feature selection techniques. Journal of Chemical Information and Modeling, Volume 53, Issue 10, Pages 2730-2742

D Newby, A. A Freitas and T. Ghafourian (2013) Coping with unbalanced class datasets in oral absorption models. Journal of Chemical Information and Modeling, Volume 53, Issue 2, Pages 461-474

T. Ghafourian, A. A Freitas and **D. Newby** (2012) The impact of training set data distribution for modelling of passive intestinal absorption. International Journal of Pharmaceutics, Volume 436, Issue 1-2, Pages 711-720

Conferences Attended

Poster presentations:

D. Newby, A.A. Freitas and T. Ghafourian, '**Comparing Multi-label Classification Methods for Provisional Biopharmaceutics Classification System (BCS) class Prediction**' at the 16th International Workshop on Quantitative Structure-Activity Relationships in Environmental and Health Sciences, 16-20th June 2014, Milan, Italy

D. Newby, A.A. Freitas and T. Ghafourian, '**Comparing Multi-label Classification Methods for Provisional Biopharmaceutics Classification System (BCS) class Prediction**' at the Spring meeting UKQSAR, 29th April 2014, Eli Lilly, Surrey

D. Newby, A.A. Freitas and T. Ghafourian, '**Pre-processing feature selection for oral absorption models**' at the Autumn meeting UKQSAR, 15th October 2013, AstraZenca, Cheshire

D. Newby, A.A. Freitas and T. Ghafourian, '**Determining permeability thresholds for oral absorption classification'** at the 5th World Conference on Drug Absorption, Transport and Delivery, 24-26th June 2013, Uppsala, Sweden

D. Newby, A.A. Freitas and T. Ghafourian, '**Comparing Caco-2 and MDCK permeability for oral absorption estimations**' at the spring meeting UKQSAR, 23rd April 2012, Unilever, Bedford (**Won best student poster**)

D. Newby, A.A. Freitas and T. Ghafourian, '**Overcoming Imbalanced Datasets in Oral Absorption Modelling**' at the Medway School of Pharmacy, 5th December 2012, (**Won best student chemistry and drug delivery poster**)

D. Newby, A.A. Freitas and T. Ghafourian, '**Overcoming Imbalanced Datasets in Oral Absorption Modelling**' at the Autumn meeting UKQSAR, 8th November 2012, Takeda, Cambridge

D. Newby, A.A. Freitas and T. Ghafourian, '**Reducing False Positives in Oral Absorption Models**' at the 15th International Workshop on Quantitative Structure-Activity Relationships in Environmental and Health Sciences, 18-22 June 2012, Tallinn, Estonia **(Won best student poster)**

D. Newby, A.A. Freitas and T. Ghafourian, '**Reducing False Positives in Oral Absorption Models**' at the spring meeting UKQSAR, 25th April 2012, Novartis, Horsham

Oral presentations:

D. Newby, A.A. Freitas and T. Ghafourian, '**Decision trees to characterise the roles of permeability and solubility on oral absorption'** at the Autumn UKQSAR meeting, 15th October 2013, AstraZenca, Cheshire **(Invited to present)**

D. Newby, A.A. Freitas and T. Ghafourian, '**Overcoming dataset bias in oral absorption modelling**' at the Medway School of Pharmacy Research Seminars, 16th January 2013, Chatham

D. Newby, A.A. Freitas and T. Ghafourian '**Reducing False Positives in Oral Absorption Models**' at the 15th International Workshop on Quantitative Structure-Activity Relationships in Environmental and Health Sciences, 18th-22nd June 2012, Tallinn, Estonia (as poster winner was invited to give an oral presentation of work)

D. Newby, A.A. Freitas and T. Ghafourian '**Computational Estimation of Intestinal Absorption**' GSK, 14th June, 2012, Stevenage