Decision trees to characterise the roles of permeability and solubility on the prediction of oral absorption

Danielle Newby[a], Alex. A. Freitas[b], Taravat Ghafourian[a,c*]

[a]Medway School of Pharmacy, Universities of Kent and Greenwich, Chatham, Kent, ME4 4TB, UK

[b]School of Computing, University of Kent, Canterbury, Kent, CT2 7NF, UK

[c] Drug Applied Research Centre and Faculty of Pharmacy, Tabriz University of Medical Sciences, Tabriz, Iran

***Corresponding Author:** Email: T.ghafourian@kent.ac.uk; Tel +44(0)1634 202952; Fax +44 (0)1634 883927

ABSTRACT

Oral absorption depends on many physiological, physiochemical and formulation factors. Two important properties that govern oral absorption are *in vitro* permeability and solubility, which are commonly used as indicators of human intestinal absorption. Despite this, the nature and exact characteristics of the relationship between these parameters are not well understood. In this study a large dataset of human intestinal absorption was collated along with *in vitro* permeability, aqueous solubility, melting point, and maximum dose for the same compounds. The dataset allowed a permeability threshold to be established objectively to predict high or low intestinal absorption. Using this permeability threshold, classification decision trees incorporating a solubility-related parameter such as experimental or predicted solubility, or the melting point based absorption potential (MPbAP), along with structural molecular descriptors were developed and validated to predict oral absorption class. The decision trees were able to determine the individual roles of permeability and solubility in oral absorption process. Poorly permeable compounds with high solubility show low intestinal absorption, whereas poorly water soluble compounds with high or low permeability may have high intestinal absorption provided that they have certain molecular characteristics such as a small polar surface or specific topology.

KEYWORDS

Intestinal absorption, permeability, solubility, decision trees, QSAR

ABBREVIATIONS USED

%HIA, percentage human intestinal absorption; BCS, Biopharmaceutics Classification System; CART, Classification and regression trees; Caco-2, Human colon adenocarcinoma cell line; FN, false negative; FP, false positive; GSE, general solubility equation; MDCK, Madin-Darby Canine Kidney; MPbAP, melting point based absorption potential; QSAR, Quantitative Structure-Activity Relationship; SE, sensitivity; SP, specificity; TN, true negative; TP, true positive

1. INTRODUCTION

The assessment of pharmacokinetic properties, especially absorption, is now well established in early drug discovery. The need to determine absorption of new chemical entities is essential for successful orally administered compounds, as well as efficacy, toxicity and other ADME (absorption, distribution, metabolism, excretion) properties [1]. The prediction of oral absorption can be carried out with experimental assays and/or the use of *in silico* models. These experimental and computer models can be used as an indication of intestinal absorption in humans, which is carried out later on in drug development. By testing drug compounds using these models, compounds with undesirable properties can be removed earlier, therefore improving cost effectiveness [2, 3].

Intestinal absorption depends on many physiological, physiochemical and formulation factors. Two important properties that govern oral absorption are permeability and solubility as utilised by the Biopharmaceutics Classification System (BCS) [4]. For a drug to be absorbed it must firstly dissolve in the gastrointestinal fluid in order to then permeate the intestinal membrane. The relationship between these properties is closely, usually inversely, related [5, 6]. As an increasing number of new chemical entities (NCE) 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 [7] .

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 [8, 9]. Early oral absorption models are too small to effectively represent all the different biological processes of absorption and the physiochemical properties including solubility [10, 11]. Most studies have removed compounds with solubility issues when modelling oral absorption [12, 13], which is not ideal due to the increasing number of poorly soluble drugs being developed. Zhao and co-workers demonstrated that predicting BCS Class II compounds (low solubility and high permeability) resulted in an overestimation of fraction absorbed by their model [12]. 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.

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 [14, 15]. The general solubility equation (GSE) is a simple method that predicts aqueous solubility using only two parameters, logP and melting point [16]. Other methods may employ more specific molecular descriptors to improve the prediction accuracy [17, 18]. GSE and its variants have been used for the estimation of oral absorption-related parameters termed absorption potential [19, 20]. 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*, and it was also found that for higher melting points absorption was limited by dose [21].

Permeability in drug discovery is routinely measured using *in vitro* cell based assays to give an indication of permeability of drug compounds in the intestine, blood brain barrier, nasal cavity and skin [22]. Apparent permeability (P_{app}) is the rate of permeation across cell monolayers and is usually measured in cm/s^{-1}. The ideal permeability model for the small intestine mimics the physical and biochemical processes of intestinal absorption [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 [23-25], which displays biological and characteristic properties of the enterocytes of the small intestine such as the brush border and tight junctions [1, 25-27]. These cells can express a variety of transporters and metabolic enzymes, allowing other transport and metabolism mechanisms to be investigated [28]. Drawbacks of this particular cell line are inter-laboratory differences, variable transporter expression, long culture time, tighter junctions compared with *in vivo* situation and lack of mucus secreting goblet cells [1, 29, 30]. 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 [31, 32]; also, the cell line HT29-MTX is a co-culture of Caco-2 cells with mucus secreting goblet cells to study the effects of mucus on absorption [33]. 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 [22, 34-36]. There are many similarities and differences between Caco-2 and MDCK cell lines. Despite this there is a linear relationship between the two shown using small compound sets [22, 34, 35].

The relationship between permeability and fraction absorption in humans can be determined numerically or categorically. From a classification perspective a permeability threshold indicates high or low intestinal absorption (absorption class). The permeability thresholds defined in the literature vary greatly 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 [13, 37-40]. For example, Artusson *et al* [37], using a dataset of 20 compounds, defined that a compound would have complete absorption if it had a permeability > 1 x10^{-6} cm/s. More recent studies have indicated higher permeability thresholds than 1 x10^{-6} to define a high absorption compound [8, 38, 41]. In a recent investigation, Varma *et al* [36], 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. Additionally, the FDA has recommended a set of high and low permeability standards with known fraction absorbed [42]. These standard compounds can be measured alongside NCEs which are then considered as highly or poorly permeable, depending on whether the permeability is greater or lower than the standards; this can then be related to fraction absorbed based on these FDA 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.

In order to see the effects of solubility and permeability on fraction absorbed, a large dataset is needed. Therefore, the first aim of this work was to expand the permeability dataset by combining data from Caco-2 and MDCK cell lines. By studying the 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. 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 of the permeability dataset collected. Using this permeability threshold, decision trees using experimental and predicted solubility and related properties such as dose number and melting point were included along with structural molecular descriptors to build classification models to predict fraction absorbed class. Therefore, the QSAR endpoint is the categorical variable indicating the ‘high’ or ‘low’ fraction absorbed class. Based on this work, 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 prediction of fraction absorbed models is explored.

2. METHODS AND MATERIALS

2.1 Datasets

With an extensive search in the literature, multiple datasets were collated consisting of data for human intestinal absorption, transport route, permeability, solubility, dose number, aqueous solubility and melting point. For each compound the name, property value, CAS number, references and additional comments from the authors relating to the data is included and can be found in the **Supporting Information I.** Whenever possible, the original literature was consulted to evaluate data quality. In some cases data from secondary sources was included when original literature could not be located.

2.1.1 Human Intestinal absorption

Intestinal absorption can be assessed and calculated from different types of data such as bioavailability, and urinary and faecal excretion mass balance studies. We used the same principles to calculate and evaluate the reliability of fraction absorbed value as defined by other works [12, 43]. Intestinal absorption values were initially obtained from the published datasets of Hou *et al* [13] and Varma *et al* [43]; this data was scrutinised by checking the original publications. An exhaustive search of the literature was then carried out and additional compounds were also added from the drug information obtained from the FDA Drugs@FDA database (accessed from June 2012 to May 2013) [44]. Where there was no numerical value defined in the literature, categorical values for fraction absorbed were also included for this dataset. At the end, the dataset consisted of 913 numerical and 19 categorical fraction absorption values creating a final dataset of 932 compounds.

2.1.2 Permeability

Apparent permeability (P_{app}) data measured in cm/s^{-1} was collected for compounds with known fraction absorption. The dataset contains apparent permeability data for the 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* [45] 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 [36, 46] 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 was used together for comparison with Caco-2.

2.1.3 Identification of absorption mechanisms

The knowledge about absorption mechanism will help with interpretation of models and give us a better understanding of the influence of transporter systems on absorption as this is increasingly important in the prediction of drug absorption. For each compound the absorption route was assessed using literature data, review articles and transporter databases. It was recorded if compounds underwent any absorption mechanism other than passive transcellular route. This included carrier mediated systems, such as efflux and influx transporters, and paracellular absorption. A total of 201 (out of 932) were identified to be absorbed via routes other than passive transcellular. It must be noted that, firstly, if no information or evidence was found to suggest alternative absorption mechanisms, this does not necessarily mean it is not a substrate of a transporter or transported via the paracellular route; it may not have been tested and/or results have not been published in the literature. Therefore, in the future we anticipate that this number could increase further when more research is carried out. Secondly, although a compound is identified as a substrate for a carrier mediated system, this does not mean that the transport system is the dominating process [47].

2.1.4 Aqueous solubility

Aqueous solubility for 483 compounds in mg/mL was obtained primarily from the AQUASOL dATAbASE (6th Edition) and SRC (PHYSPROP) databases (http://esc.srcinc.com/fatepointer/search.asp) and the literature. Solubility was converted to log molar units (M) and log mg/mL units in this work. 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.

In addition to these values, predicted solubility values were also utilised and compared with experimental in the modelling section of this work. Solubility was calculated by the revised general solubility equation (GSE) using experimental melting point and calculated logP.[16] (**Equation 1** below).

$$\text{Log Sol (GSE)} = 0.5 - 0.01\,(MP - 25) - logP \qquad \textbf{(1)}$$

2.1.5 Dose number

Dose number is a dimensionless number used to determine high or low solubility in the Biopharmaceutical Classification System (BCS) [4]. It is calculated using the solubility and maximum strength dose (**Equation 2**).

$$D_o = (M_o / V_o) / S \qquad \textbf{(2)}$$

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 for the compounds in this dataset from the British National Formulary (2012)[48], FDA electronic orange book 2012 (accessed December 2012-January 2013) and Martindale (2009) [49]. Where there were still missing values, an extensive literature search was carried out and the values presented are the authors' 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.

2.1.6 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.

2.1.7 Melting 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 [21]. As shown by **Equation 3** below.

$$\text{MPbAP} = 0.5 - 0.01\,(MP - 25) - \log(4 * Max\ Dose) \qquad \textbf{(3)}$$

2.2 Calculated molecular descriptors

Calculated 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.

2.3 Training 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 78 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 are shown in **Table 1**.

Table 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

2.4 Classification and regression trees (CARTs)

STATISTICA v11 (StatSoft Ltd.) software was used for permeability threshold determination and classification of compounds using CART analysis. CARTs (called C&RT in the STATISTICA software) use decision trees to solve regression and classification problems developed by Breinman *et al* [50]. Hence, in this work the QSAR models are represented as decision trees (a type of graph). According to the observed %HIA values in the data set, compounds were placed into either the "high" class if %HIA was equal to or greater than a specified HIA cut-off (e.g.50%) or the "low" class if %HIA was less than this specified %HIA cutoff. In this work binary classification of (low or high HIA) was carried out using calculated molecular descriptors from the chemical structure, permeability and solubility related parameters. The QSAR models (in the form of decision trees) used in this work were validated by measuring the predictive accuracy of model predictions (prediction of "High" or "Low" oral absorption class) for the compounds in the validation set, as described earlier (section 2.3 – training and validation sets).

Preliminary results indicated that permeability and not solubility was the dominant property selected statistically by CART. Therefore in order to gauge the relative importance of these two parameters, the decision trees were built in two phases. The first phase forced CART to select a suitable permeability threshold for different HIA class definitions. The second phase involved forcing CART to choose thresholds for solubility and related parameters for the second split of the decision tree. After this, CART was allowed to build the remainder of the tree automatically using structural molecular descriptors. These trees were compared with a CART tree developed using the parameters selected automatically by the tree from permeability or solubility parameters or the molecular descriptors provided.

2.5 Permeability threshold determination using CART

The permeability threshold is the numerical value chosen by CART that best predicts HIA class. In this work 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 CART analysis was restricted to only one split to give the permeability threshold. This threshold was tested using a validation set of 78 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 bias. Based on previous works the use of misclassification costs has shown improved model accuracy [51]. 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 data set. 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.

2.6 Permeability and 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 lower HIA class definition such as 30-70% due to the lower number of poorly absorbed compounds only poor models could be achieved. 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 work 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 CART 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 then second split variable (note that CART still chooses the cut-off point automatically) and then the tree was allowed to grow automatically. Stopping factors were used to prevent overfitting of the CART trees and was the minimum number of compounds for splitting. This was set at 11 for the permeability only CART trees and eight for permeability and solubility trees.

2.7 Statistical significance of the models

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 was the squared correlation coefficient, r^2 forced through the origin. For correlation analysis the Pearson's correlation coefficient 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 [52]. 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 CART in this work was measured using sensitivity (SE), specificity (SP) and SP × SE. Sensitivity is the ratio of correct classifications for the highly absorbed compound class (SE = TP/(TP + FN)), where TP is the number of true positives and FN is the number of false negatives. Specificity is the ratio of correct classifications of poorly absorbed compounds (SP = TN/(TN + FP)), where TN is the number of true negatives and FP is the number of false positives. In this work overall accuracy is defined by specificity multiplied by sensitivity (SP × SE). This measure represents the overall predictive performance of both high and low class prediction. In addition, this measure will not be overly influenced by the classification accuracy of the majority high absorption class, and it has been used in previous investigations [51, 53].

3. RESULTS AND DISCUSSIONS

In this work 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. **Table 2** shows the collated data which is available in the **supporting information I**, where n denotes the number of compounds for each property. This data was used in order to develop models for predicting high/low oral absorption and to explore suitability of different solubility and permeability measures from different sources as descriptors of intestinal absorption.

Table 2. Data sets collated from the literature

Property	n
Human intestinal absorption	932
Caco-2 permeability	386
MDCK permeability	246
Aqueous solubility	482
Dose number	465
Melting point	609

In terms of permeability, we 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. Other cell lines such as MDCK, 2/4/A1 and HT29-MTX have also been used to assess compound permeability. There have been a few studies, which show the linear relationship between these cell lines. For example, Braun *et al* [22] studied the relationship between Caco-2 and MDCK cell lines and from 14 compounds achieved an r^2 of 0.86. However, Adveef *et al* [35] achieved a r^2 of 0.90 using a dataset of 79 compounds.

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 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 1**) where the transport routes have also been highlighted. Out of 185 compounds in this figure, 96 compounds were found to be substrates of a transporter system and 11 compounds have been suggested to be absorbed to some extent via paracellular route.

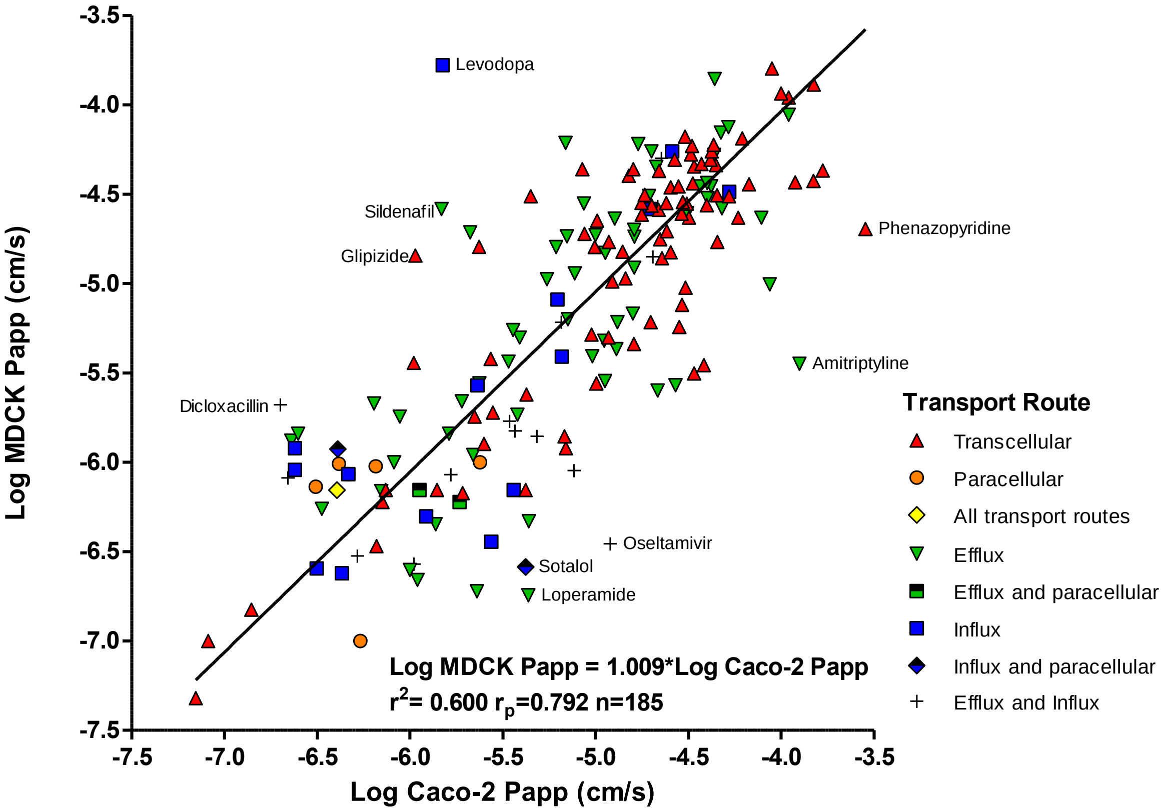

Figure 1. Linear relationship between Caco-2 and MDCK apparent permeability for 185 compounds

It can be seen in the plot that Caco-2 and MDCK permeability of majority of compounds regardless of their absorption routes correlate well with each other. However, there are compounds that deviate significantly from this line and removal of 9 outlier compounds (compound names shown in the **Figure 1**) improves the correlation significantly (**Table 3**). Details of the outlier compounds and a description of reasons can be found in **Supporting Information II**. A better linear relationship between the two cell lines is also achieved when only compounds undergoing passive transcellular absorption are plotted (**Table 3**). It may be noted in **Table 3** that the correlation 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 classed in Class II of Biopharmaceutics classification system (BCS) [54, 55].

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* [34] compared the apparently 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. We crossed referenced the remaining compounds used by Irvine with our database and identified an additional 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 in comparing the two cell lines is not suitable. For the P-gp substrates highlighted in the work, it was stated the reason they could not be identified was due to saturation of the transport mechanism in the assay. Braun *et al* [22] 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 3. Statistical parameters for the linear relationship between MDCK and Caco-2 permeability measured using PRISM

Datasets	**r^2 (with intercept)**	**r^2 (non-intercept)**	**R_p**	**R_s**
All compounds (185)	0.63	0.60	0.79	0.79
Passive transcellular (83)	0.71	0.69	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

We have complied a table that compares the cells and small intestine in terms of species origin, tightness of the cell junctions and also the transporter and enzyme expressions (**Table S1 in Supporting Information III**). One thing to note is the lack of information/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, we tried to accommodate an overview of expression from the human small intestine [56]. It can be seen from **Table S1** 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 [57, 58], 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 [30, 59]. A recent investigation has found no evidence of CYP3A4 expression in MDCK II cells [60]. 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 physiochemical properties of the compound [61]. 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 [54, 55] (**Figure 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 4).**

Table 4. 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 and paracellular (H)	1	ranitidine

From **Table 4**, 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.

We 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 4**. 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 (**Figures S1-S7 and Table S2** in the **Supporting Information III**). 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 S1 in Supporting Information III**). 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) showed no significant difference between Caco-2 and MDCK permeabilities ($p > 0.05$). In addition paired t tests for all different absorption mechanism groups and no significant differences between the two cell lines for these absorption groups were found. Therefore, we 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.

3.2 Determining permeability threshold for an effective oral absorption

In this work we use the large dataset of combined Caco-2 and MDCK permeability and a statistical method (CART) to identify statistically valid permeability threshold for high/low oral absorption. Using CART 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 CART models using different misclassification cost ratios for false positives: false negatives (FP:FN) were generated [51, 53]. The results below show the permeability threshold selected by the CART analyses and the accuracy, specificity and sensitivity of the class prediction (**Table 5**).

Table 5. The permeability thresholds selected by CART and HIA class prediction with equal and higher misclassification costs applied to false positives when high HIA defined as higher than 30, 50, 70, 80 and 90%

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

t: training set; v: validation set

It can be seen in **Table 5** that using high ratios of (FP:FN) misclassification costs results in 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 is not successful at all (**Model 1 Table 5**) but increasing the cost of false positives to five times that of the false negatives results in a high accuracy of classification and a robust threshold of -5.98 (in log units) (model 6). It must be noted here that different high/low definitions of HIA result 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 [51, 53]. Therefore by applying higher misclassification costs to reduce false positives, this has shifted the permeability threshold in order to reduce the number of false positives due to the under representation of the poorly absorbed class (**Figure 2**). 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 CART when applying higher misclassification costs from **Table 5** can be shown below (**Figure 2**) when plotting fraction absorbed against permeability for the training and validation sets.

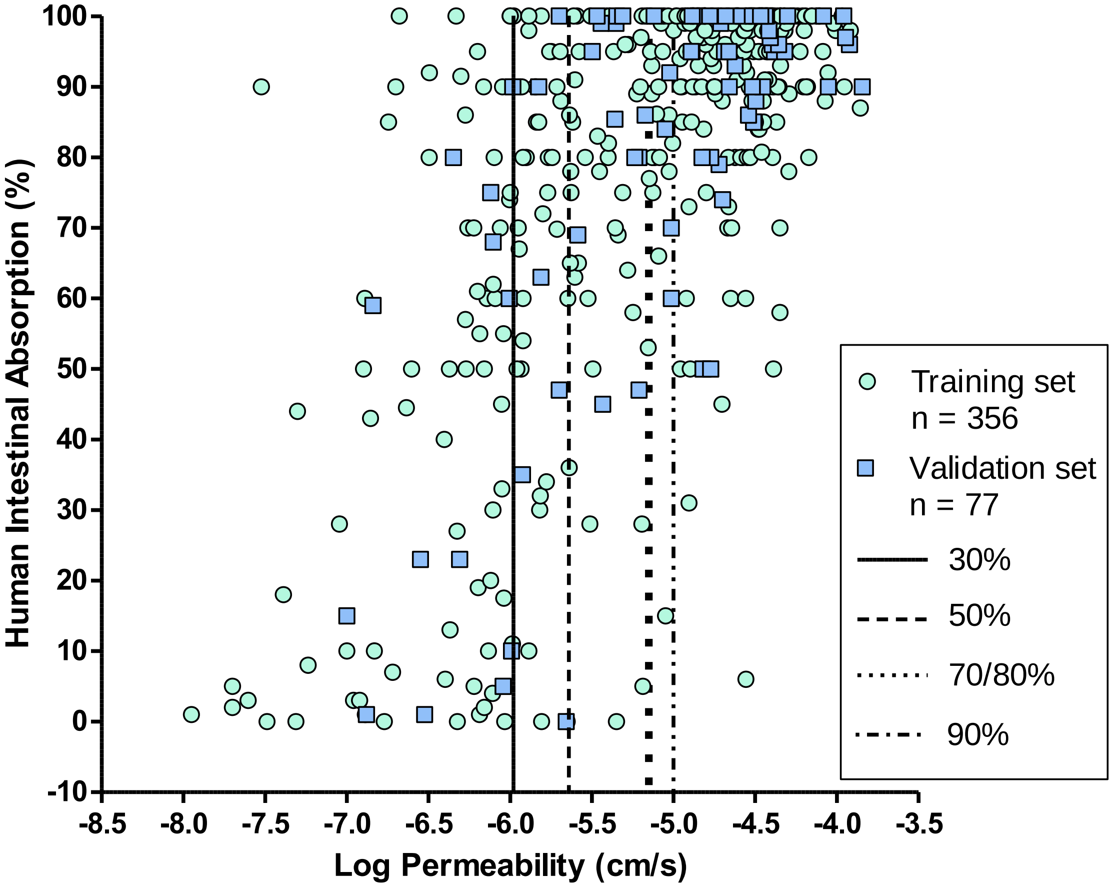

Figure 2. Permeability thresholds determined by CART analysis with higher misclassification costs applied to false positives for different HIA cut offs of 30%, 50%, 70%, 80% and 90% on %HIA versus permeability plot including areas of outliers (A= low permeability, high oral absorption; B = high permeability, low oral absorption)

As can be seen by **Figure 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 [32, 36, 62]. However, there are too few points at the lower plateau region to justify fitting a sigmoidal fit from statistical point of view; in spite of this we 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 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 the figure (**Figure 2**) using boxes A and B. Compounds with low permeability but high fraction absorbed (Region A on **Figure 2**) have been identified as mainly highly soluble and substrates for influx carrier mediated transporters. Examples of these are ribavirin and lamivudine [63, 64]. 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 from this dataset (Region B on **Figure 2**). Examples of compounds in this outlier group are lovastatin and tacrolimus [65, 66].

Although the liver is the main metabolising organ, gut metabolism can contribute significantly to overall metabolism and should be considered [67]. 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 are both highly absorbed and highly permeable. Reasons for why some compounds are susceptible to gut metabolism and others are not even though they are both CYP3A4 substrates could be due to the different biotransformation rate by this enzyme, solubility/ dissolution rate, permeation rate, dose amount and substrate affinity [67-69]. A list of these compounds in regions A and B in **Figure 2** can be found in the **Supporting Information II**.

3.3 Oral absorption prediction using solubility, dose number and melting point

From **Figure 2,** we 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 misclassifications 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 maybe poorly soluble compounds or those undergoing gut metabolism.

CART 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 material and methods section. The permeability for ≥ 80% absorption (at -5.15 log scale according to **Table 5**) was used to develop the models. The 80% class definition was chosen as when using lower HIA% values to define high or low absorption led to 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 our 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 CART models produced for the prediction of HIA class (HIA> or ≤80%) using permeability and solubility related parameters and molecular descriptors are shown in **Table 6**. 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, CART picks the most significant parameter out of all the molecular descriptors and physicochemical properties available. In **Table 6**, in model 1 after permeability as the first split variable, CART automatically builds the rest of the tree by selecting the most significant property/molecular descriptor. For models 2-4, solubility; 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 CART 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 6. The results of CART 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	Accuracy (SP x SE)	Sensitivity (SE)	Specificity (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	0.72	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	0.820
					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 CART 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

From **Table 6** 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 6** 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 for a good prediction model (Model 6 in **Table 6**). Dose number is used to define high and low solubility for the BCS system [4, 42]. 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 6** incorporate solubility and predicted solubility especially for highly permeable compounds. Unlike GSE solubility which was used on both sides of the CART trees, MPbAP only yielded good models when used for splitting on the high permeability compounds. Experimental solubility in two units, mg/ml or molar, have been used in models. Solubility in M, which takes into account the molecular weight and is smaller for high molecular weight compounds, was 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 Supporting Information III**). 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 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, particles size, etc) which could mask previous solubility limiting effects of such compounds.

The top molecular descriptors used in models 1-12 in **Table 6** are polar surface area (PSA) and Balaban topological index. Both of these descriptors are related to both absorption and solubility prediction models [70, 71]. PSA is the area of the van der Waals surface that arises from oxygen and nitrogen atoms or hydrogen atoms bound to these atoms [70]. The Balaban topological index, *J*, is the average-distance sum connectivity and relates to the shape of the molecule [72]. The next popular descriptors are sum of absolute charges on each atom of the molecule (ABSQ) and lowest unoccupied molecular orbital energy (LUMO) calculated by VAMP [73].

3.4 Selected CART models

In order to generally compare models 1-12 from **Table 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. When the dataset is large, e.g. in the case of model 1, there are more compounds that maybe harder to classify 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 it may be more suitable for generalization ability for new compounds, as it covers a wider chemical space compared with model 8. Moreover the only experimental parameter used in this model is melting point that is used for the calculation of MPbAP. We 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 CART models are presented in **Figures 3-5**.

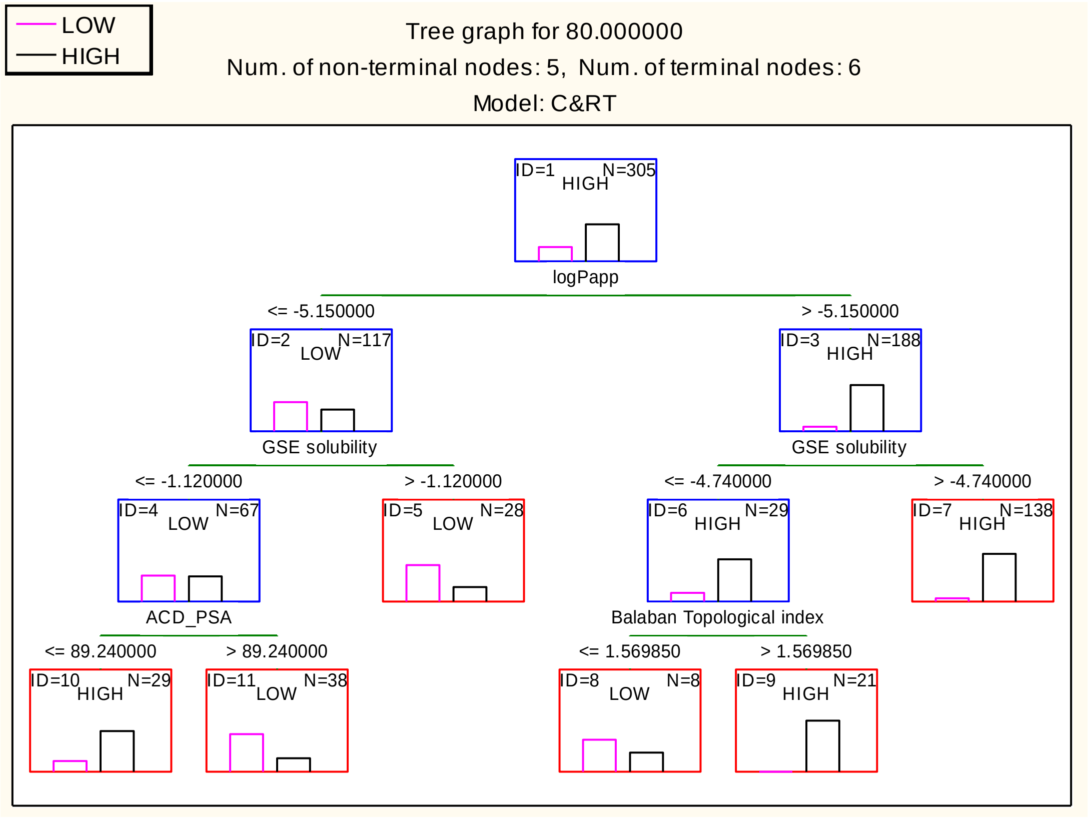

Figure 3. Model 3 CART permeability and predicted solubility (GSE) model when higher misclassification costs of two to reduce false positives were applied to low GSE solubility node

In **Figure 3**, Model 3, permeability is used as the first CART split variable and then calculated solubility from GSE equation on both sides of the tree was used as the second split variable. Polar surface area and Balaban index were picked automatically by the CART analysis. The model shows that highly permeable and highly soluble compounds have high intestinal absorption (node 7). Moreover, compounds with low predicted solubility (≤-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 3; ziprasidone and tiagabine are misclassified as poorly absorbed when in fact they have HIA ≥ 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 [72]. The relationship between Balaban index and solubility with reference to melting point has been shown previously in the literature [15]. In spite of this there is not much difference between the calculated GSE solubilities between the two nodes although there is a significant difference betwee 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 are highly absorbed only for compounds with predicted solubility ≤ -1.12 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, therefore classified as highly-absorbed according to this tree. The reasons for misclassifications is mostly due to efflux mechanisms reducing the absorption of compounds and examples include nadolol and norfloxacin which both have low PSA and classed as highly absorbed but are observed to have poor oral absorption due to transporter effects [31, 74]. 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 presence of more such compounds in this dataset, may have led to further split of this node based on solubility to class compounds with extremely low aqueous solubility as poorly soluble.

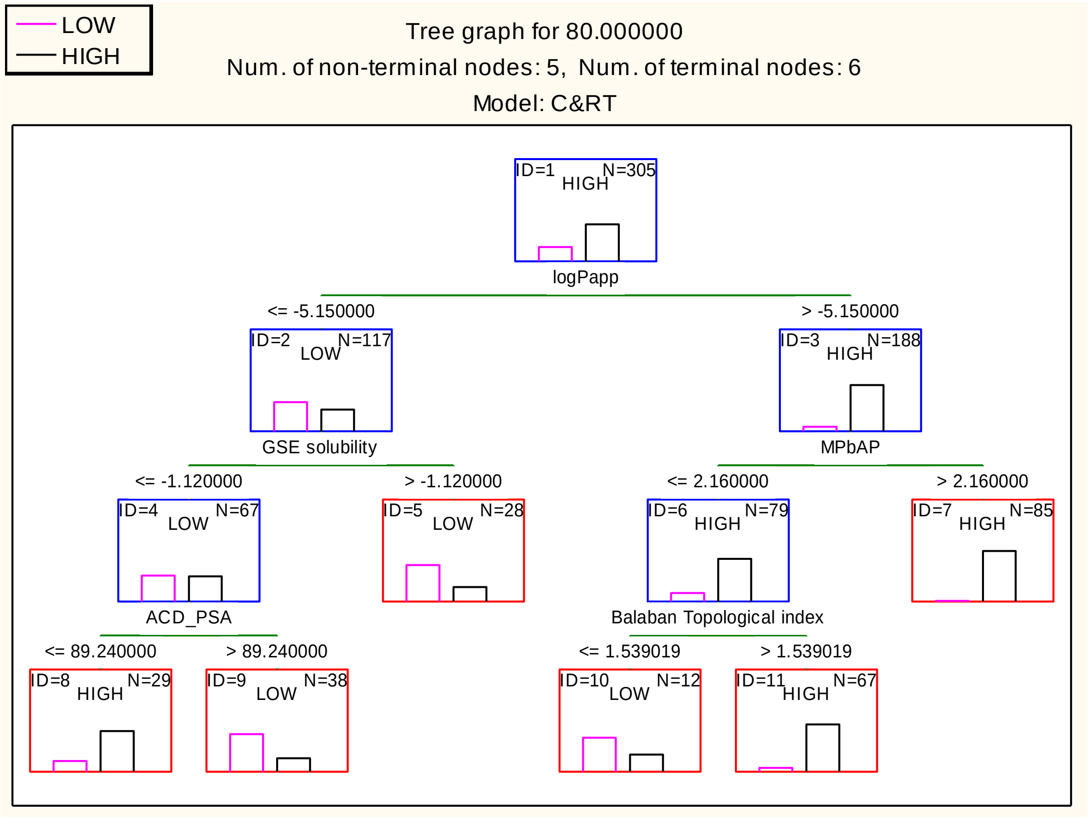

Figure 4. Model 7 CART permeability, predicted solubility (GSE) and MPbAP model when higher misclassification costs of two to reduce false positives were applied to GSE node

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 set. The descriptors used in this tree are the same as in **Figure 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.

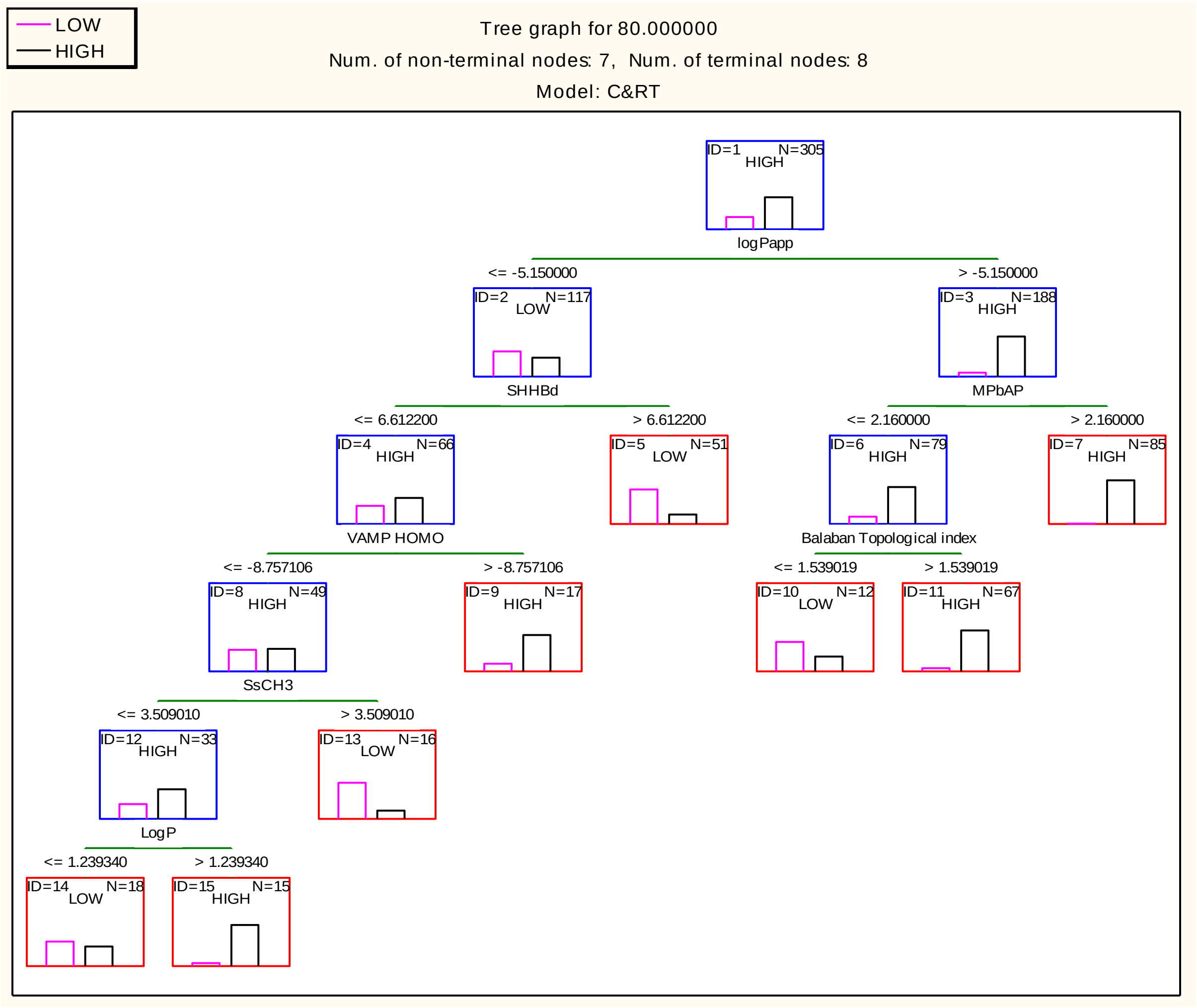

Figure 5. Model 12 CART permeability and MPbAP model when higher misclassification costs of two to reduce false positives were applied to permeability node

From **Figure 5** classification of highly permeable compounds in node 3 is the same as Figure 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, where compounds are likely to be poorly absorbed if two or more of the following rules are broken: more than > 5 hydrogen bond donors, > 10 hydrogen bond acceptors, logP > 5 and molecular weight > 500Da [75]. 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.

3.5 Discussion of related literature

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 in **Table 7.**

Table 7. 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 (1991) [37]	Caco-2	>1 ≤ 0.1	100 < 1	20
Yee (1997) [41]	Caco-2	< 1 1-10 >10	0-20 20-70 70-100	35
Bergstrom (2003) [9]	Caco-2	≤ 0.2 ≥ 1.6	≤ 20 ≥ 80	27
Hou (2007) [13]	Caco-2	≥ 6 .0	High (>80)	69
Di (2011) [40]	MDCK II	~ 3	Low/medium (<80) High (>80)	19
Varma (2012) [36]	MDCKII*	≥ 5.0	≥ 80/90	97
Pham-The (2013) [62]	Caco-2	~ 0.7 ≥ 16 .0	< 30 ≥ 85	324

*MDCKII strain (MDCK-LE) cell line with isolated low endogenous efflux transporter expression

Early permeability thresholds defined by works in the literature are based on small compound datasets. Artusson *et al* [37] set a permeability threshold of > 1 x 10^{-6} cm/s for complete absorption based on 20 compounds. Based on other works in the literature this value is too low to predict complete absorption, where other works have permeability thresholds one order of magnitude higher. For example, from **Table 7**, Yee *et al* [41] has stated > 10 x 10^{-6} cm/s 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 [13]. In comparison, our permeability thresholds are statistically defined by CART rather than a subjective determination; the thresholds picked by CART 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* [13] of 6 x 10^{-6} cm/s is based on data from numerous sources and is very similar to our 70 - 90% class permeability thresholds.

Di *et al* (2011) [40] used MDCK II cells with low efflux endogenous transporter expression (MDCK-LE) to define a threshold of 3 x 10^{-6} cm/s to distinguish between low/medium absorbed compounds (< 80% HIA) and highly absorbed compounds. A dataset published by Varma *et al* (2012) [36] using the MDCK-LE cell line shows that the permeability threshold defined ROC analysis using this cell line (≥ 5.0 x 10^{-6} cm/s) is similar to Caco-2 thresholds in the literature, and this value is in agreement with CART 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) [62] 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 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.

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 [6]. In spite of this the main reasons for poor oral absorption have been shown to be either poor permeability or poor solubility or both [76].

The results of this work indicate that permeability is the most important parameter influencing oral absorption prediction. Permeability was always picked as top molecular descriptor when building CART 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 [4, 12, 77]. 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 [78].

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 study, 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 in log unit, which is quite high when comparing with the threshold values suggested in the literature for BCS classes II and IV [4]. 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 648 compounds[13]. 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 data sets 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 [10, 79]. 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 [13, 80]. This simplifies the resulting models by removing those compounds with these 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 solubility and dose dependency was defined and not used in the majority of the initial models. However upon inclusion of these compounds with solubility issues the resulting models had higher error [81]. It was also noted however that the more insoluble a compound the lower the resulting absorption. In a later study compounds identified with no solubility issues were used to built 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 [12]. Our oral absorption models are able to predict oral absorption class even with poor solubility for majority of compounds 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* (2001) [12], 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 [82]. 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 our results. They also found various contour plots that incorporating solubility improves classification of HIA based on permeability data by about 10%; therefore showing that potentially using solubility in models is 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 our results, predicted solubility 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.

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 comparing a linear relationship between these two cell lines, even for compounds with different absorption mechanisms.

Secondly, using the combined permeability dataset, a permeability threshold for various levels of oral absorption was investigated using CART analysis. Due to the larger number of highly absorbed compounds, misclassification costs were applied and improved the threshold definition statistically. The thresholds obtained from the objective CART 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 CART 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 oral absorption. These molecular descriptors were shown to be necessary for oral absorption models 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.

ASSOCIATED CONTENT

Supporting information I contains the dataset of 932 compounds with HIA%, Caco-2 permeability, MDCK permeability, aqueous solubility, melting point and the references. Supporting information II contains compound lists and information regarding the outliers in Figures 1 and 3 including references. Supporting Information III contains a table comparing the differences in transporter and enzyme expression between the human small intestine, Caco-2 and MDCK cell lines, the significance testing and graphs for the different absorption mechanisms when comparing Caco-2 and MDCK cell lines, all the models (CART decision trees) produced from this work, and finally a list of molecular descriptor utilised in the 12 models presented in this work.

REFERENCES

[1] D.A. Volpe, Variability in Caco-2 and MDCK cell-based intestinal permeability assays, J. Pharm. Sci., 97 (2008) 712-725.

[2] A. Boobis, U. Gundert-Remy, P. Kremers, P. Macheras, O. Pelkonen, In silico prediction of ADME and pharmacokinetics - Report of an expert meeting organised by COST B15, Eur. J. Pharm. Sci., 17 (2002) 183-193.

[3] H. van de Waterbeemd, E. Gifford, ADMET in silico modelling: towards prediction paradise?, Nat. Rev. Drug Discov., 2 (2003) 192-204.

[4] G.L. Amidon, H. Lennernas, V.P. Shah, J.R. Crison, A theoretical basis for a biopharmaceutic drug classification - The correlation of in-vitro drug product dissolution and in-vivo bioavalability, Pharm. Res., 12 (1995) 413-420.

[5] S.T. Buckley, S.M. Fischer, G. Fricker, M. Brandl, In vitro models to evaluate the permeability of poorly soluble drug entities: Challenges and perspectives, Eur. J. Pharm. Sci., 45 (2012) 235-250.

[6] C.A. Lipinski, Drug-like properties and the causes of poor solubility and poor permeability, J. Pharmacol. Toxicol. Methods, 44 (2000) 235-249.

[7] J.M. Miller, A. Beig, B.J. Krieg, R.A. Carr, T.B. Borchardt, G.E. Amidon, G.L. Amidon, A. Dahan, The Solubility-Permeability Interplay: Mechanistic Modeling and Predictive Application of the Impact of Micellar Solubilization on Intestinal Permeation, Mol. Pharm., 8 (2011) 1848-1856.

[8] V. Pade, S. Stavchansky, Link between drug absorption solubility and permeability measurements in Caco-2 cells, J. Pharm. Sci., 87 (1998) 1604-1607.

[9] C.A.S. Bergstrom, M. Strafford, L. Lazorova, A. Avdeef, K. Luthman, P. Artursson, Absorption classification of oral drugs based on molecular surface properties, J. Med. Chem., 46 (2003) 558-570.

[10] M.D. Wessel, P.C. Jurs, J.W. Tolan, S.M. Muskal, Prediction of human intestinal absorption of drug compounds from molecular structure, J. Chem. Inf. Comp. Sci., 38 (1998) 726-735.

[11] T. Niwa, Using general regression and probabilistic neural networks to predict human intestinal absorption with topological descriptors derived from two-dimensional chemical structures, J. Chem. Inf. Comp. Sci., 43 (2003) 113-119.

[12] Y.H. Zhao, M.H. Abraham, J. Le, A. Hersey, C.N. Luscombe, G. Beck, B. Sherborne, I. Cooper, Rate-limited steps of human oral absorption and QSAR studies, Pharm. Res, 19 (2002) 1446-1457.

[13] T.J. Hou, J.M. Wang, W. Zhang, X.J. Xu, ADME evaluation in drug discovery. 7. Prediction of oral absorption by correlation and classification, J. Chem. Inf. Mod., 47 (2007) 208-218.

[14] J.M. Wang, T.J. Hou, Recent Advances on Aqueous Solubility Prediction, Comb. Chem. High Throughput Screen., 14 (2011) 328-338.

[15] T. Ghafourian, A.H.A. Bozorgi, Estimation of drug solubility in water, PEG 400 and their binary mixtures using the molecular structures of solutes, Eur. J. Pharm. Sci., 40 (2010) 430-440.

[16] N. Jain, S.H. Yalkowsky, Estimation of the aqueous solubility I: Application to organic nonelectrolytes, J. Pharm. Sci., 90 (2001) 234-252.

[17] A.L. Cheng, K.M. Merz, Prediction of aqueous solubility of a diverse set of compounds using quantitative structure-property relationships, J. Med. Chem., 46 (2003) 3572-3580.

[18] X.Q. Chen, S.J. Cho, Y. Li, S. Venkatesh, Prediction of aqueous solubility of organic compounds using a quantitative structure-property relationship, J. Pharm. Sci., 91 (2002) 1838-1852.

[19] J.B. Dressman, G.L. Amidon, D. Fleisher, Absorption potential: Estimating the fraction absorbed for orally administered compounds, J. Pharm. Sci., 74 (1985) 588-589.

[20] T. Sanghvi, N. Ni, S.H. Yalkowsky, A simple modified absorption potential, Pharm. Res, 18 (2001) 1794-1796.

[21] K.A. Chu, S.H. Yalkowsky, An interesting relationship between drug absorption and melting point, Int. J. Pharm., 373 (2009) 24-40.

[22] A. Braun, S. Hammerle, K. Suda, B. Rothen-Rutishauser, M. Gunthert, S.D. Kramer, H. Wunderli-Allenspach, Cell cultures as tools in biopharmacy, Eur. J. Pharm. Sci., 11 (2000) S51-S60.

[23] P.V. Balimane, S.H. Chong, R.A. Morrison, Current methodologies used for evaluation of intestinal permeability and absorption, J. Pharmacol. Toxicol. Methods, 44 (2000) 301-312.

[24] J. Fogh, G. Trempe, Human Tumor Cells In Vitro in: J. Fogh (Ed.) Human Tumor Cells In Vitro, Plenum Press, New York, 1975, pp. 115-141.

[25] I.J. Hidalgo, T.J. Raub, R.T. Borchardt, Characterization of the human colon carcinoma cell line (Caco-2) as a model system for intestinal epithelial permeability, Gastroenterology, 96 (1989) 736-739.

[26] P. Artursson, Epithelial transport of drugs in cell-culture.1. A model for studying the passive diffusion of drugs over intestinal absorptive (Caco-2) cells, J. Pharm. Sci., 79 (1990) 476-482.

[27] M. Pinto, S. Robineleon, M.D. Appay, M. Kedinger, N. Triadou, E. Dussaulx, B. Lacroix, P. Simonassmann, K. Haffen, J. Fogh, A. Zweibaum, Enterocyte-like differentiation and polarization of the human-colon carcinoma cell-line caco-2 in culture, Biol. Cell, 47 (1983) 323-330.

[28] R.B. van Breemen, L. Li, Caco-2 cell permeability assays to measure drug absorption, Expert Opin. Drug Metab. Toxicol., 1 (2005) 175-185

[29] M.J. BriskeAnderson, J.W. Finley, S.M. Newman, The influence of culture time and passage number on the morphological and physiological development of Caco-2 cells, Proc. Soc. Exp. Biol. and Med., 214 (1997) 248-257.

[30] E. Le Ferrec, C. Chesne, P. Artusson, D. Brayden, G. Fabre, P. Gires, F. Guillou, M. Rousset, W. Rubas, M.L. Scarino, In vitro models of the intestinal barrier - The report and recommendations of ECVAM Workshop 46, ATLA-Altern. Lab. Anim., 29 (2001) 649-668.

[31] P. Matsson, C.A.S. Bergstrom, N. Nagahara, S. Tavelin, U. Norinder, P. Artursson, Exploring the role of different drug transport routes in permeability screening, J. Med. Chem., 48 (2005) 604-613.

[32] S. Tavelin, J. Taipalensuu, L. Soderberg, R. Morrison, S.H. Chong, P. Artursson, Prediction of the oral absorption of low-permeability drugs using small intestine-like 2/4/A1 cell monolayers, Pharm. Res, 20 (2003) 397-405.

[33] C. Hilgendorf, H. Spahn-Langguth, C.G. Regardh, E. Lipka, G.L. Amidon, P. Langguth, Caco-2 versus Caco-2/HT29-MTX co-cultured cell lines: Permeabilities via diffusion, inside- and outside-directed carrier-mediated transport, J. Pharm. Sci., 89 (2000) 63-75.

[34] J.D. Irvine, L. Takahashi, K. Lockhart, J. Cheong, J.W. Tolan, H.E. Selick, J.R. Grove, MDCK (Madin-Darby canine kidney) cells: A tool for membrane permeability screening, J. Pharm. Sci., 88 (1999) 28-33.

[35] A. Avdeef, K.Y. Tam, How Well Can the Caco-2/Madin-Darby Canine Kidney Models Predict Effective Human Jejunal Permeability?, J. Med. Chem., 53 (2010) 3566-3584.

[36] M.V. Varma, I. Gardner, S.J. Steyn, P. Nkansah, C.J. Rotter, C. Whitney-Pickett, H. Zhang, L. Di, M. Cram, K.S. Fenner, A.F. El-Kattan, pH-Dependent Solubility and Permeability Criteria for Provisional Biopharmaceutics Classification (BCS and BDDCS) in Early Drug Discovery, Mol. Pharm., 9 (2012) 1199-1212.

[37] P. Artursson, J. Karlsson, Correlation between oral-drug absorption in humans and apparent drug perrmeability coefficients in human intestinal epithelial (Caco-2) cells, Biochem. Biophys. Res. Commun., 175 (1991) 880-885.

[38] P. Stenberg, U. Norinder, K. Luthman, P. Artursson, Experimental and computational screening models for the prediction of intestinal drug absorption, J. Med. Chem., 44 (2001) 1927-1937.

[39] M.C. Gres, B. Julian, M. Bourrie, V. Meunier, C. Roques, M. Berger, X. Boulenc, Y. Berger, G. Fabre, 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, Pharm. Res, 15 (1998) 726-733.

[40] L. Di, C. Whitney-Pickett, J.P. Umland, H. Zhang, X. Zhang, D.F. Gebhard, Y.R. Lai, J.J. Federico, R.E. Davidson, R. Smith, E.L. Reyner, C. Lee, B. Feng, C. Rotter, M.V. Varma, S. Kempshall, K. Fenner, A.F. El-Kattan, T.E. Liston, M.D. Troutman, Development of a New Permeability Assay Using Low-Efflux MDCKII Cells, J. Pharm. Sci., 100 (2011) 4974-4985.

[41] S.Y. Yee, In vitro permeability across Caco3 cells (colonic) can predict in vivo (small intestinal) absorption in man - Fact or myth, Pharm. Res, 14 (1997) 763-766.

[42] CDER/FDA, Waiver of In Vivo Bioavailability and Bioequivalence Studies for Immediate-Release Solid Oral Dosage Forms Based on a Biopharmaceutics Classification System, in: U.S. Department of Health and Human Services - Center for Drug Evaluation and Research. Guidance for Industry, 2000.

[43] M.V.S. Varma, R.S. Obach, C. Rotter, H.R. Miller, G. Chang, S.J. Steyn, A. El-Kattan, M.D. Troutman, Physicochemical Space for Optimum Oral Bioavailability: Contribution of Human Intestinal Absorption and First-Pass Elimination, J. Med. Chem., 53 (2010) 1098-1108.

[44] FDA, FDA approved drug products Drugs@FDA (http://www.accessdata.fda.gov/scripts/cder/drugsatfda/index.cfm) (accessed from June 2012 to May 2013)

[45] H.P. The, I. Gonzalez-Alvarez, M. Bermejo, V.M. Sanjuan, I. Centelles, T.M. Garrigues, M.A. Cabrera-Perez, In Silico Prediction of Caco-2 Cell Permeability by a Classification QSAR Approach, Mol. Inform., 30 (2011) 376-385.

[46] M.V.S. Varma, K. Sateesh, R. Panchagnula, Functional role of P-glycoprotein in limiting intestinal absorption of drugs: Contribution of passive permeability to P-glycoprotein mediated efflux transport, Mol. Pharm., 2 (2005) 12-21.

[47] K. Sugano, M. Kansy, P. Artursson, A. Avdeef, S. Bendels, L. Di, G.F. Ecker, B. Faller, H. Fischer, G. Gerebtzoff, H. Lennernaes, F. Senner, Coexistence of passive and carrier-mediated processes in drug transport, Nat. Rev. Drug Discov., 9 (2010) 597-614.

[48] BNF, British National Formulary, September 64 ed., BMJ Group and Pharmaceutical Press, London, 2012.

[49] Martindale The Complete Drug Reference, 36 ed., Pharmaceutical Press, London, 2009.

[50] L. Breiman, J. Friedman, C.J. Stone, R.A. Olshen, Classification and Regression Trees, 1 ed., Chapman and Hall/CRC, Boca Raton, 1984.

[51] D. Newby, A.A. Freitas, T. Ghafourian, Coping with Unbalanced Class Data Sets in Oral Absorption Models, J. Chem. Inf. Mod., 53 (2013) 461-474.

[52] G.J. Hahn, Fitting Regression Models with No Intercept Term, J. Qual. Technol., 9 (1977) 56-61.

[53] D. Newby, A.A. Freitas, T. Ghafourian, Pre-processing Feature Selection for Improved C&RT Models for Oral Absorption, J. Chem. Inf. Mod., 53 (2013) 2730-2742.

[54] Z. Gao, Development of a Continuous Dissolution/Absorption System—a Technical Note, AAPS PharmSciTech, 13 (2012) 1287-1292.

[55] A. Mehramizi, B. Alijani, M. Pourfarzib, F.A. Dorkoosh, M. Rafiee – Tehrani, Solid Carriers for Improved Solubility of Glipizide in Osmotically Controlled Oral Drug Delivery System, Drug. Dev. Ind. Pharm., 33 (2007) 812-823.

[56] G. Englund, F. Rorsman, A. Ronnblom, U. Karlbom, L. Lazorova, J. Grasjo, A. Kindmark, P. Artursson, Regional levels of drug transporters along the human intestinal tract: Co-expression of ABC and SLC transporters and comparison with Caco-2 cells, European J. Pharm. Sci., 29 (2006) 269-277.

[57] A.-L.B. Ungell, Caco-2 replace or refine?, Drug. Discov. Today: Technol., 1 (2004) 423-430.

[58] C. Hilgendorf, G. Ahlin, A. Seithel, P. Artursson, A.L. Ungell, J. Karlsson, Expression of thirty-six drug transporter genes in human intestine, liver, kidney, and organotypic cell lines, Drug Metab. Dispos., 35 (2007) 1333-1340.

[59] R. Hayeshi, C. Hilgendorf, P. Artursson, P. Augustijns, B. Brodin, P. Dehertogh, K. Fisher, L. Fossati, E. Hovenkamp, T. Korjamo, C. Masungi, N. Maubon, R. Mols, A. Mullertz, J. Monkkonen,

C. O'Driscoll, H.M. Oppers-Tiemissen, E.G.E. Ragnarsson, M. Rooseboom, A.L. Ungell, Comparison of drug transporter gene expression and functionality in Caco-2 cells from 10 different laboratories, Eur. J. Pharm. Sci., 35 (2008) 383-396.

[60] Y. Quan, Y. Jin, T.N. Faria, C.A. C. A. Tilford, A. He, D.A. Wall, R.L. Smith, B.S. Vig, Expression Profile of Drug and Nutrient Absorption Related Genes in Madin-Darby Canine Kidney (MDCK) Cells Grown under Differentiation Conditions, Pharmaceutics, 4 (2012) 314-333.

[61] P. Shah, V. Jogani, T. Bagchi, A. Misra, Role of Caco-2 cell monolayers in prediction of intestinal drug absorption, Biotechnol. Progr., 22 (2006) 186-198.

[62] H. Pham-The, I. González-Álvarez, M. Bermejo, T. Garrigues, H. Le-Thi-Thu, M.Á. Cabrera-Pérez, The Use of Rule-Based and QSPR Approaches in ADME Profiling: A Case Study on Caco-2 Permeability, Mol. Inform., 32 (2013) 459-479.

[63] S. Shugarts, L.Z. Benet, The Role of Transporters in the Pharmacokinetics of Orally Administered Drugs, Pharm. Res, 26 (2009) 2039-2054.

[64] G. Minuesa, C. Volk, M. Molina-Arcas, V. Gorboulev, I. Erkizia, P. Arndt, B. Clotet, M. Pastor-Anglada, H. Koepsell, J. Martinez-Picado, 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, J. Pharmacol. Exp. Ther., 329 (2009) 1187-1187.

[65] M.F. Hebert, Contributions of hepatic and intestinal metabolism and P-glycoprotein to cyclosporine and tacrolimus oral drug delivery, Adv. Drug Deliv. Rev., 27 (1997) 201-214.

[66] W. Jacobsen, G. Kirchner, K. Hallensleben, L. Mancinelli, M. Deters, I. Hackbarth, K. Baner, L.Z. Benet, K.F. Sewing, U. Christians, Small intestinal metabolism of the 3-hydroxy-3-methylglutary-coenzyme A reductase inhibitor lovastatin and comparison with pravastatin, J. Pharmacol. Exp. Ther., 291 (1999) 131-139.

[67] M. Gertz, A. Harrison, J.B. Houston, A. Galetin, Prediction of Human Intestinal First-Pass Metabolism of 25 CYP3A Substrates from In Vitro Clearance and Permeability Data, Drug Metab. Dispos., 38 (2010) 1147-1158.

[68] U. Fagerholm, Prediction of human pharmacokinetics - gut-wall metabolism, J. Pharm. Pharmacol., 59 (2007) 1335-1343.

[69] J.H. Lin, M. Chiba, T.A. Baillie, Is the role of the small intestine in first-pass metabolism overemphasized?, Pharmacol. Rev., 51 (1999) 135-157.

[70] D.E. Clark, Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 1. Prediction of intestinal absorption, J. Pharm. Sci., 88 (1999) 807-814.

[71] C.A. Bergstrom, In silico predictions of drug solubility and permeability: two rate-limiting barriers to oral drug absorption, Basic Clin Pharmacol Toxicol, 96 (2005) 156-161.

[72] A.T. Balaban, Highly discriminating distance-based topological index, Chem. Phys. Lett., 89 (1982) 399-404.

[73] J. Gasteiger, M. Marsili, Iterative partial equalization of orbital electronegativity - a rapid access to atomic charges, Tetrahedron, 36 (1980) 3219-3228.

[74] G. Merino, A.I. Alvarez, M.M. Pulido, A.J. Molina, A.H. Schinkel, J.G. Prieto, Breast cancer resistance protein (BCRP/ABCG2) transports fluoroquinolone antibiotics and affects their oral availability, pharmacokinetics, and milk secretion, Drug Metab. Dispos., 34 (2006) 690-695.

[75] C.A. Lipinski, F. Lombardo, B.W. Dominy, P.J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, Adv. Drug Deliv. Rev., 23 (1997) 3-25.

[76] K.T. Savjani, A.K. Gajjar, J.K. Savjani, Drug solubility: importance and enhancement techniques, ISRN Pharm, (2012) Article ID 195727, 1-10.

[77] K. Sugano, Fraction of a dose absorbed estimation for structurally diverse low solubility compounds, Int. J. Pharm., 405 (2011) 79-89.

[78] S. Stegemann, F. Leveiller, D. Franchi, H. de Jong, H. Linden, When poor solubility becomes an issue: From early stage to proof of concept, Eur. J. Pharm. Sci., 31 (2007) 249-261.

[79] J.P.F. Bai, A. Utis, G. Crippen, H.D. He, V. Fischer, R. Tullman, H.Q. Yin, C.P. Hsu, L. Jiang, K.K. Hwang, Use of classification regression tree in predicting oral absorption in humans, J. Chem. Inf. Comp. Sci., 44 (2004) 2061-2069.

[80] W.J. Egan, K.M. Merz, J.J. Baldwin, Prediction of drug absorption using multivariate statistics, J. Med. Chem., 43 (2000) 3867-3877.

[81] Y.H. Zhao, J. Le, M.H. Abraham, A. Hersey, P.J. Eddershaw, C.N. Luscombe, D. Boutina, G. Beck, B. Sherborne, I. Cooper, J.A. Platts, Evaluation of human intestinal absorption data and subsequent derivation of a quantitative structure-activity relationship (QSAR) with the Abraham descriptors, J. Pharm. Sci., 90 (2001) 749-784.

[82] H. Pham-The, T. Garrigues, M. Bermejo, I. González-Álvarez, M.C. Monteagudo, M.Á. Cabrera-Pérez, Provisional Classification and in Silico Study of Biopharmaceutical System Based on Caco-2 Cell Permeability and Dose Number, Mol Pharm, 10 (2013) 2445-2461.

CAPTIONS (in order of appearance in manuscript)

Table 1. Compound numbers used in the training and validation sets for decision tree analysis

Table 2. Data sets collated from the literature

Figure 1. Linear relationship between Caco-2 and MDCK apparent permeability for 185 compounds

Table 3. Statistical parameters for the linear relationship between MDCK and Caco-2 permeability measured using PRISM

Table 4. The different identified absorption mechanism of the 185 compounds

Table 5. The permeability thresholds selected by CART and HIA class prediction with equal and higher misclassification costs applied to false positives when high HIA defined as higher than 30, 50, 70, 80 and 90%

Figure 2. Permeability thresholds determined by CART 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 outliers (A= low permeability, high oral absorption; B = high permeability, low oral absorption)

Table 6. The results of CART analysis for the best permeability and solubility related trees using permeability threshold for ≥ 80% or < 80% HIA as the first split

Figure 3. Model 3 CART permeability and predicted solubility (GSE) model when higher misclassification costs of two to reduce false positives were applied to low GSE solubility node

Figure 4. Model 7 CART permeability, predicted solubility (GSE) and MPbAP model when higher misclassification costs of two to reduce false positives were applied to GSE node

Figure 5. Model 12 CART permeability and MPbAP model when higher misclassification costs of two to reduce false positives were applied to permeability node

Table 7. Examples of permeability thresholds determined by the literature