MODELLING CELLULAR PERMEABILITY VIA CARRIER MEDIATED TRANSPORT

By

Tendai Ishmael Sarupinda

November 2015

A THESIS SUBMITTED TO THE UNIVERSITY OF HERTFORDSHIRE

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE BY RESEARCH

Abstract

The relative importance of passive diffusion and carrier mediated transport processes to membrane permeability of drugs is a subject of current debate. Passive diffusion and carrier mediated transport are the two main methods by which drugs permeate the cell membrane. The permeability of molecules through membranes can have an impact on their absorption, distribution, metabolism and excretion (ADME) properties. It is therefore important to be able to predict the extent to which novel molecules can permeate the cell membrane. *In vitro* models of human intestinal absorption can be used to predict the likelihood of molecules permeating the human intestinal epithelium. Quantitative structure activity relationships (QSAR) techniques explain the relationship between molecular structure and cellular permeability. Current QSAR methods make use of physicochemical and structural property descriptors. These descriptors are able to predict the membrane permeability of molecules via passive diffusion rather than via membrane transporters. The aim of this study was to develop novel descriptors of carrier mediated transport that can be used in the development of QSAR models of permeability. The concept of metabolite likeness was investigated for its utility as a measure of the likelihood of molecules undergoing carrier mediated transport. This investigation found that approved drugs are generally more similar to human endogenous metabolites than molecules found in commercial databases. The use of a protein target prediction tool, PIDGIN, was also investigated. This study found that a relatively small number of membrane transporters that are expressed in caco-2 cells have models available in PIDGIN. New QSAR models of membrane permeability were developed using physicochemical and structural property descriptors and in combination with the novel descriptors of carrier mediated transport. Novel models for predicting drug efflux ratio were developed and perform well in validation tests. Comparisons of predictive performance between QSAR models generated from physicochemical property descriptors alone and in combination with 'carrier-mediated transport descriptors' were carried out. The general observation was that the novel descriptors of carrier mediated transport pursued did not significantly improve the predictive performance of models. However, some substructures from

the MACCS keys list, which are relevant to protein binding, were found to be important determinants of caco-2 permeability of molecules and could potentially be used to identify molecules that may undergo active transport. The performance of logistic regression classification models of efflux ratio was 88%. Not many studies have developed QSAR models of efflux ratio. This is a relatively novel approach which could be useful in identifying, and thus help to avoid, potential substrates of efflux transporters in drug discovery.

Acknowledgements

I would like to thank my industrial placement supervisors, Dr. Tim James and Dr. Mike Bodkin from Evotec, and my academic supervisor Professor Mire Zloh from the University of Hertfordshire for their combined support throughout my project. I have learnt a great deal from these people and without their guidance, I would not have been able to complete this project.

To the Research Informatics group at Evotec, you have allowed me to grow as a person. Many thanks for making me feel welcome and part of the group, and of course for the wonderful experience ☺

Last but certainly not least, I would like to thank my family and loved ones. Without your constant support and encouragement, I would not be where I am today.

Table of Contents

List of Figures

List of Tables

1. INTRODUCTION

Not only do successful drugs have to demonstrate potency against their intended protein targets, but they must also possess good absorption, distribution, metabolism and excretion (ADME) properties. As of 2004, 50% of drug candidates failed due to poor ADME and toxicity properties in drug development (Hou, Zhang, Xia, Qiao, & Xu, 2004). Due to these high attrition rates, ADME properties are being considered in the early stages of drug discovery and development, resulting in significant reduction of ADME related attrition. However, the cost associated with experimental evaluation (*in-vitro or in-vivo*) of ADME properties is great. Therefore, cheaper alternatives such as computational (*in-silico*) techniques for predicting ADME properties prior to synthesis are considered (Cheng, Li, et al., 2012).

Drug bioavailability is dependent on the extent to which the drug is absorbed into cells. Absorption is therefore an important property, the prediction of which is of immense value in drug discovery. While many factors influence the extent of absorption and therefore bioavailability, membrane permeability is arguably a very important factor. This is evidenced by the number of studies aiming to predict membrane permeability through in-silico, in vitro and in-vivo methods (Deli, Ábrahám, Kataoka, & Niwa, 2005; Dolghih & Jacobson, 2013a; T. Hou, Wang, Zhang, Wang, & Xu, 2006; Sevin et al., 2013; Stenberg, Norinder, Luthman, & Artursson, 2001).

To be effective, most drugs must cross cell membranes to reach, and interact with, their intracellular biological targets. The structure of cell membranes is such that they are effective regulators of passage of substances into and out of the cell. Depending on the chemical properties of the substance, passage into the cell can occur by one (or more) of four methods: passive transcellular diffusion, carrier mediated transport, transcytosis and paracellular transport (through intercellular gaps between cells). Passive diffusion generally occurs when sufficiently lipophilic molecules interact with phospholipids in the membrane (Sugano et al., 2010). There is no direct evidence suggesting that passive diffusion is the main method of drug

permeation, however, it is often the case in drug discovery projects that molecules are designed to be more lipophilic in the hope of making the compound more permeable via passive diffusion. This approach is understandable given the difficulty of designing molecules that can interact with specific membrane transporters as well as their protein targets. To design such a drug, one would require the drug to possess chemical features necessary for binding to the membrane transporter as well as to the intended target protein. Lipinski's rule of 5 is a widely known rule used in drug discovery for assessing whether a molecule is likely to be orally active in humans (Lipinski, Lombardo, Dominy, & Feeney, 2012). However, the rule does not apply to substrates of membrane transporters and is therefore not applicable to all molecules.

The increase in computational power has accompanied an increase in *in-silico* methods in drug discovery (Ekins, Mestres, & Testa, 2007). A wide range of *in-silico* methods have been developed, ranging from complex mathematical simulations of biological systems using molecular and quantum mechanical (QM) methods (van der Kamp & Mulholland, 2013) to statistical methods of predicting the behaviour of molecules in physiological environments. Quantitative structure-activity relationships (QSAR) modelling is an example of a statistical method often applied to predict bioactivity or other properties of interest (e.g. permeability) of molecules as a function of numerical molecular descriptors (Dehmer, Varmuza, & Bonchev, 2012). Some advantages of QSAR modelling include the speed by which models can be generated and the fact that molecular descriptors required for generating the models can be calculated from the chemical structure alone (Yee & Wei, 2012).

One limitation of current QSAR methods of predicting membrane permeability is that they make use of physicochemical property descriptors which are better at predicting the permeability of molecules via passive diffusion than carrier-mediated transport. Because cell membranes contain several membrane transporters, it is likely that the permeability of drug molecules is greatly influenced by carrier mediated transport. Both membrane transporters and drugs are promiscuous in their nature; a drug can bind to several transporters and transporters can recognise several drugs which may or may not be structurally similar (Kell, Dobson, Bilsland, &

Oliver, 2013). This makes it a difficult task to develop predictive models for identifying potential substrates of membrane transporters. Unsurprisingly, not many studies of QSAR incorporate descriptors of carrier mediated transporters. The aim of this study is to assess the utility of the recently proposed concept of metabolite-likeness and predictions of affinity to membrane transporters as novel descriptors of carrier mediated transport for use in the development of predictive QSAR models of membrane permeability.

1. 1 Structure of the Cell Membrane

To address the permeability of molecules through cell membranes, it is important to consider the relationship between the structure of cell membranes and how this relates to its function in terms of passage of substances.

The cell membrane is made up of a phospholipid bilayer in which membrane proteins and carbohydrates etc. can be found embedded (Goñi, 2014). Cell membranes are often referred to as biological barriers which regulate the entry and exit of substances into and out of the cell respectively, maintaining an environment under which the cell exhibits optimal functioning (Sugano et al., 2010). The phospholipids that make up the bilayer are made up of polar phosphate groups more commonly known as 'hydrophilic heads' facing the aqueous media, and fatty acid chains that make up the hydrophobic tails. A common representation of the cell membrane is the fluid mosaic model (Figure 1) (Nicolson, 2014).

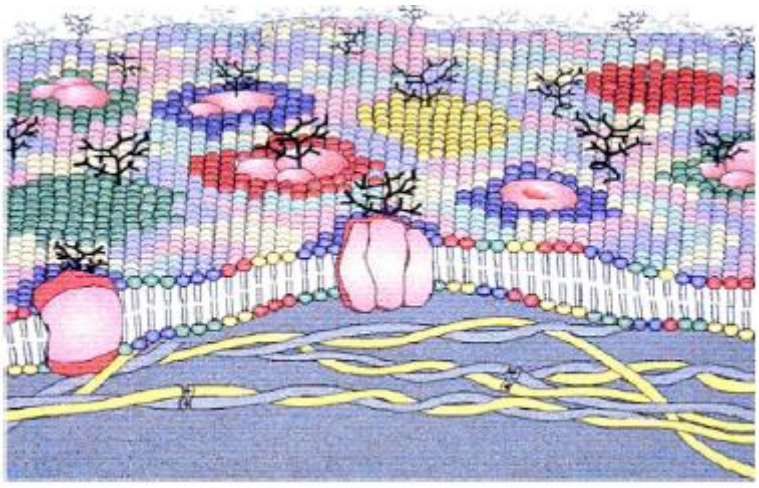

Figure 1: Illustration of the fluid mosaic model (Nicolson, 2014)

Intracellular organelles, such as mitochondria and nuclei, may also have membranes that regulate the passage of substances (Stryer & Gumport, 1995). Consequently, drugs that target proteins inside these organelles require properties that enable them to cross both the cellular membrane and the organelle membrane. Cell membranes also contain receptors on the extracellular side to which specific external stimuli can bind and trigger a reaction cascade or signal transduction within that particular cell (Rothfield, 1971). They also contain a wide range of membrane transport proteins that facilitate the movement of molecules across the membrane (Goñi, 2014). Membrane transporters that can be found in the membrane include ion channels, and uptake and efflux transporters. Uptake and efflux transporters are involved in carrier-mediated transport of molecules into and out of the cells respectively. Energy, in the form of ATP, is required for uptake and efflux pumps. Ion channels facilitate passive diffusion of ions, such as sodium and potassium ions, that cannot permeate through the lipid bilayer due to their charge.

It is generally thought that many drug molecules are transported across biological membranes via passive transcellular diffusion (through the membrane lipids) at a rate related to their lipophilicity. However, the chemical features considered important for molecules to interact with membrane lipids are also important for interaction with membrane transporters. A recent hypothesis proposes that carrier mediated transport accounts for the majority of membrane drug transport in biological systems (Dobson & Kell, 2008a; Kell, Dobson, & Oliver, 2011).

The structure of the membrane is therefore closely related to its function. There are several methods by which molecules can enter the cell. In order to address cellular permeability of molecules, one must take into account all the possible methods.

1. 2 Methods of membrane transport

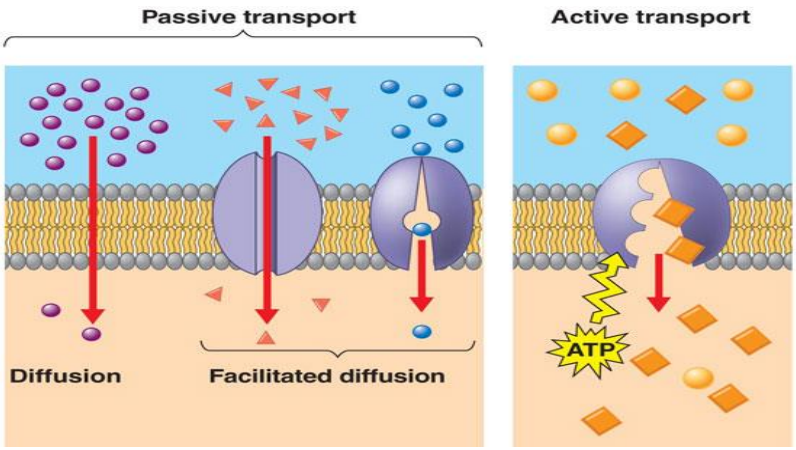

Figure 2: Methods of transport across cell membranes

1.2.1 Passive diffusion of drugs through cell membranes

Movement of molecules across the membrane through lipids is referred to as transcellular passive diffusion (Figure 2). No energy is required for this process and molecules with a higher degree of lipophilicity tend to cross the membrane more readily along their concentration gradient (Stryer & Gumport, 1995). It is widely believed that passive diffusion across cell membranes is the predominant transport method of most drugs (Kerns & Di, 2008). The process of passive diffusion is described in detail by Kerns and Di (Kerns & Di, 2008). The process begins with the shedding of water molecules surrounding the drug molecule where hydrogen bonds are broken, allowing the molecule to pass through the polar regions of the phospholipids. Molecules then move to the tightly packed lipid chains surrounding the glycerol backbone and then move towards the more disordered aliphatic chains in the middle of the bilayer. Small molecules pass through the tightly packed chains more readily whilst highly lipophilic molecules are more permeable in the non-polar centre of the membrane. Molecules then pass through the second layer of the membrane and are rehydrated by water molecules, forming hydrogen bonds.

1.2.2 Carrier-mediated transport of drug molecules

Movement of drug molecules across cell membranes can be mediated by membrane transporters (Figure 2). These membrane transporters may be responsible for uptake of molecules into the cell (uptake transporters) or efflux out of the cell (efflux

transporters). Examples of uptake transporters include the organic uptake transport proteins (OATPs) which consist of 11 members (Roth, Obaidat, & Hagenbuch, 2012) and the monocarboxylate transporters (MCTs) responsible for transport of molecules that contain a single carboxylate group (Halestrap & Wilson, 2012). Examples of efflux transporters include the widely studied Permeability glycoprotein (P-gp) and the breast cancer resistance protein (BCRP). The degree to which membrane transporters contribute to the permeation depends on the structure of the molecule and the cell membrane in question (Sugano et al., 2010). This process is often referred to as active transport because energy, in the form of adenosine triphosphate (ATP), is required. Membrane transporters can therefore affect the ADME properties of drug molecules and substrate binding to membrane transporters should be considered in membrane permeability studies.

1.3 Relative importance of passive diffusion and carrier mediated transport

The relative importance of passive diffusion and active transport to the pharmacokinetic profile of drug molecules is a topic of current debate (Sugano et al., 2010). The general perception is that carrier-mediated transport is rare compared with passive diffusion. The most likely reason for this perception is because molecules that are more lipophilic tend to exhibit higher permeability compared with less lipophilic molecules (Di et al., 2012; Lipinski et al., 2012; Sugano et al., 2010).

However, a review of the literature suggests no evidence supporting this perception exists. Recently, Kell et al. (Dobson & Kell, 2008a; Kell, 2015a; Kell et al., 2011) proposed that carrier-mediated transport is in fact the main method of transport and passive diffusion is negligible, a proposal that has, unsurprisingly, received much criticism. Di et al (Di et al., 2012; Kerns & Di, 2008) and Sugano et al. (Sugano et al., 2010), among others, propose the co-existence of passive and carrier mediated transport in drug permeation.

The notion of passive diffusion being the main method of permeation has been challenged. Dobson and Kell (2008) have proposed that carrier-mediated transport is the 'rule rather than the exception' (Dobson & Kell, 2008b). If that is the case, successful *in-silico* modelling of carrier-mediated transport processes should improve the accuracy of *in-silico* permeability predictions. Models of permeability are currently based on physicochemical descriptors that are more capable of predicting the likelihood of a molecule undergoing passive diffusion than carrier-mediated transport. The proposal that drug cell permeation is carrier-mediated only has triggered a debate on the matter (Di et al., 2012; Dobson & Kell, 2008b; Kell, 2015b; Kell et al., 2011).

According to Kell et al., there is increasing evidence pointing at the idea that drugs enter cells by 'hitchhiking on carriers' that normally transport nutrients and endogenous metabolites across the membrane (Kell et al., 2011). One of the arguments raised in support of a carrier-mediated only hypothesis is that permeation through lipid membranes is negligible because the protein-lipid mass ratio in cell membranes ranges between 1:1 and 3:1. Should this be the case, one would expect carrier-mediated transport to be more prevalent than passive diffusion. While this ratio may seem to support the notion of a carrier-mediated only process, it is worth analysing the abundance of proteins that are membrane transporters rather than just proteins because the cell membrane is known to contain various types of proteins, many of which are surface receptors. If the protein-lipid ratios mentioned above are in fact correct, it is reasonable to conclude that artificial membranes, such as the parallel artificial membrane permeability assay (PAMPA) (Bermejo et al., 2004), do not resemble biological membranes because they are made up entirely of lipids. The suggestion is that these assays are irrelevant and therefore cannot be used as an indicator of permeability in biological membranes. Indeed the above study mentions cases where molecules have 100-fold greater permeability in cell systems than in artificial membranes. The most likely reason for this is the influence of carrier-mediated transport. However, other reports have shown that some compounds are more permeable in artificial membranes than in cell-based systems. A recent report claims that permeability of a small set of

carboxylic acids obeys Overton's rule which states that more lipophilic molecules cross the phospholipid bilayer quicker (Li, Hu, & Malmstadt, 2011). For that set of molecules, one would expect passive diffusion to be the predominant transport mechanism.

Some molecules tested in different cell-lines (e.g. caco-2 and MDCK) show different permeability values (Di et al., 2012). The most likely reason for these differences is the difference in morphology of the relevant cell lines. If passive diffusion is the main method of permeation for all drug molecules, one may expect fairly equal permeability values between cell lines. However, other factors may influence differences in permeability between cell lines e.g. difference in thickness and the degree of heterogeneity of the membranes (Sugano et al., 2010). It is the major differences in permeability that can therefore be attributed to the difference in membrane transporters expressed in different cell lines. For example if a molecule has high permeability in a caco-2 cell line but is not permeable in a MDCK cell, one would expect carrier-mediated transport to be the main factor and passive diffusion to be negligible for that particular molecule.

When passive diffusion is dominant, transport rates are expected to increase linearly with concentration and rates are expected to be equal in both directions (Matsson et al., 2015). However, concentration dependence of transport rate does not exclude active transport. At high concentrations, membrane transporters become saturated and transport rate reaches a 'steady state' and becomes independent of drug concentration. At low drug concentration, saturation of membrane transporters may not be observed and one can expect such linear relationships between concentration and transport rate.

Collectively, the evidence suggests that both passive diffusion and carrier-mediated transport can contribute to the permeability of molecules and the relative importance of each process depends on the molecule and the cell membrane in question.

1.3.1 Physicochemical properties affecting permeability

There are many factors known to influence the permeability of molecules including hydrogen bonding capability, lipophilicity, molecular size and ionization state (Hou et al., 2006).

The octanol-water partition coefficient (K_{ow}) is widely used in drug discovery to measure the lipophilicity (or hydrophobicity) of a molecule (Waring, 2010). Lipophilicity is an important descriptor that measures the extent to which a molecule partitions in aqueous or hydrophobic environments such as the phospholipid section of cell membranes. Due to its widespread use, lipophilicity is considered one of the most successful physicochemical property descriptors in medicinal chemistry (Testa, Crivori, Reist, & Carrupt, 2000). Many studies have found correlations between lipophilicity and bioactivity or other behaviour of interest such as cellular permeability (Kah & Brown, 2008; Waring, 2010). The partition coefficient, logP, quantifies the partitioning of neutral molecules between octanol and water. The disadvantage of logP is that it is only applicable to neutral molecules. Many molecules are ionisable and the logP value of ionisable compounds is pH dependent. The distribution ratio, (logD), is used to measure the lipophilicity of both neutral and ionisable compounds and is therefore more useful than logP in the drug development process (Kah & Brown, 2008). Molecules that are highly lipophilic are generally more permeable than molecules that are less lipophilic (Arnott & Planey, 2012). However, molecules that are too lipophilic have low solubility therefore the relationship between lipophilicity and permeability is non-linear. This is one of the reasons why passive diffusion is regarded the main process of drug cellular permeation. The basis of this reason is that the higher the lipophilicity, the more likely it is that the molecule will interact with the phospholipid regions of the cell membrane (Arnott & Planey, 2012; Kah & Brown, 2008; Waring, 2010).
However, high lipophilicity does not render a molecule unable to interact with membrane transporters. Chemical features that contribute to high lipophilicity are also relevant to protein binding e.g. a large number of methyl groups increase lipophilicity but these are also important for hydrophobic interactions with proteins

(Kell & Oliver, 2014). This is one of the points raised by Kell et al. who suggested drug transport is “essentially carrier mediated only” (Kell et al., 2011).

For orally administered drugs to permeate the intestinal membrane, they must be soluble. Solubility is a measure of amount or concentration of the drug that can be dissolved in a particular medium. Poor solubility values often result in poor ADME properties. There is usually an inverse relationship between solubility and lipophilicity therefore a trade-off between the two properties is often required when developing drugs with optimal ADME properties.

The majority of molecules can exist in both ionized and unionised states because they are either weak bases or acids. The amount of ionized and unionized drug depends on the pH of the solvent. The unionized form of a drug is widely reported to permeate the cell membrane via passive diffusion. It was also perceived that the unionized form of the drug was impermeable but studies suggest this is not always the case (Kah & Brown, 2008). It is likely that the unionized form permeates via passive diffusion while the ionized form permeates via carrier-mediated transport. The acid dissociation constant, pKa, is used as a measure of the extent of ionization at a given pH value and is dependent on the strength of acidic and basic groups in the molecule (Manallack, 2011).

The significance of molecular size in permeability is demonstrated by its inclusion in the Lipinski’s rule of 5 which states that, among other rules, a molecule is unlikely to be absorbed if it has molecular weight greater than 500 Daltons (Lipinski et al., 2012; Newby, 2014). The increase in size of molecules is accompanied with poor permeability in particular via passive diffusion. There are however exceptions for molecules that undergo carrier-mediated transport; molecules of large size have been reported to permeate via carrier-mediated transport (H. Sun & Pang, 2007).

1.4 *In-vitro* models of cell permeability

The advancement in combinatorial chemistry and high throughput techniques has caused an increase in numbers of molecules in compound libraries that require

experimental testing (Irvine et al., 1999). ADME properties of drug molecules influence the amount of drug molecules that reach circulation and consequently their targets. It is therefore essential to have models for predicting the likelihood of compounds being permeable early in the drug discovery pipeline. Animal models can be used for such tests; however, the cost of animal model studies makes this approach unattractive for early drug discovery purposes. In vitro methods are a cheaper and faster approach and consequently, a wide range of in vitro systems have been developed to predict human intestinal absorption. In vitro models range from simple methods that determine the partitioning of a molecule in different solvents to more complex cell culture and whole tissue methods (Deli et al., 2005).

1.4.1 Colon adenocarcinoma (Caco-2) cells

The most commonly used in vitro method for predicting absorption is the caco-2 cell (Calcagno, Ludwig, Fostel, Gottesman, & Ambudkar, 2006; Guangli & Yiyu, 2006; Pham-The et al., 2011). While derived from the human colon, caco-2 cells differentiate to resemble the intestinal epithelium when cultured under specific conditions. Because they retain many morphological and functional properties of the intestinal epithelium, they are considered the gold standard for human intestinal permeability studies. Caco-2 permeability is widely reported to correlate well with human intestinal absorption (Kerns & Di, 2008). The widely acknowledged importance of caco-2 cells for permeability studies has caused an upsurge in computational studies aiming to predict the caco-2 permeability of compounds.

1.4.2 Parallel artificial membrane permeability assay (PAMPA)

PAMPA is a method used to determine the permeability of compounds through an artificial membrane made up of lipids (Bermejo et al., 2004). The main use of the PAMPA assay is to assess the likelihood and the extent to which molecules undergo passive diffusion. One obvious disadvantage of PAMPA is that only passive diffusion can be predicted. While passive diffusion is regarded to be the main method by which drugs enter cells, the main outcome of the recent debate on the relative importance of passive diffusion and active transport was the agreement of a coexistence of both processes. This presents another obvious limitation of PAMPA: a lack of carrier-mediated transport processes. The results obtained from such cell

lines have been questioned. If carrier-mediated transport is indeed more prevalent over passive diffusion, permeability through PAMPA may be irrelevant. However, many have observed good correlation between PAMPA permeability and *in-vivo* permeability.

1.4.3 Mardin-Darby Canine Kidney (MDCK)

The (MDCK) epithelial cells can also be used in permeability screening of chemical compounds (Irvine et al., 1999). While the cell line is derived from canine (dog) kidney, reports suggest that MDCK permeability show good correlation with both Caco-2 and human intestinal permeability. Because of the time required to culture Caco-2 cells, the MDCK cell line is sometimes used in early stage drug discovery to predict membrane permeability (Volpe, 2008). The obvious limitations lie in the differences in morphology between canine kidney and human intestinal epithelial cells. Another limitation of MDCK cell lines is the variation in observed permeability values of the same molecules between different laboratories (Volpe, 2008). This variation is most likely due to the different culture conditions applied in different laboratories which may cause differences in the morphology of the cell lines.

1.5 The concept of molecular similarity

The concept of molecular similarity has widespread use in drug discovery and medicinal chemistry. The similar property principle states that structurally similar molecules are likely to share similar properties (Eckert & Bajorath, 2007; Mestres & Maggiora, 2006). This principle forms the basis of the concept of molecular similarity.

The concept of bioisosterism is a typical example of an application of molecular similarity (Patani & LaVoie, 1996). Bioisosteric groups are functional groups that can be interchanged within a molecular structure and still maintain activity. This approach has been employed, in particular, to replace toxic groups with safer functional groups (Patani & LaVoie, 1996).

The structural similarity of ligands has been applied as a method of relating protein function. In this approach, ligands of each protein are collected and based on their overall similarity proteins can be considered similar or dissimilar in terms of function. A study by Keiser et al. (Keiser et al., 2007) illustrates how protein functioning can be predicted based on the similarity of their ligands. The aim was to group and relate proteins based on the similarities of their ligands and to create a similarity map of enzymes and receptors. Depending on the availability of data, it may be possible for a similar approach to be taken to identify transporter substrates and to also relate membrane transporters with respect to similarity of their substrates.

Other methods based on the concept of molecular similarity include the development of target prediction tools. Known active molecules against a particular target are collected and based on the similarity coefficients, new molecules can be predicted as substrates or non-substrates of the relevant target. An example is the PIDGIN tool that makes use of activity and inactivity information for a particular target (Mervin et al., 2015). The tool contains predictive models for 1080 protein targets for which novel molecules can be predicted as substrates or non-substrates depending on structural similarity to active or inactive molecules for the relevant target.

There are however some limitations to the concept of molecular similarity. One limitation is that not all structurally similar molecules have similar biological activity profiles as single point changes in molecular structure may cause drastic changes in biological activity (Guha & Van Drie, 2008; Maggiora, 2006). Pairs of molecules with small differences in structure but significant differences in activity are often referred to as activity cliffs. While activity cliffs demonstrate that molecular similarity does not always correspond with similar biological activity, studies of single point changes can identify important sites for biological activity (Stumpfe, Hu, Dimova, & Bajorath, 2014). Another limitation is that molecular similarity is highly context dependent (Mestres & Maggiora, 2006). For example, two molecules may have similar molecular weights but their lipophilicity may differ drastically.

1.5.1 Molecular Fingerprints

The process of comparing molecules requires methods of representing the molecular structure and metrics for assessing structural similarity. A wide range of descriptors are available for use in calculating molecular similarity. Descriptors are designed to represent molecular composition in a given descriptor space. The first stage in structural similarity comparisons is the generation of molecular descriptors. Once descriptors are generated, a measure of similarity is required. An example is the widely used Tanimoto coefficient which calculates similarity based on binary representations of molecules (Cereto-Massagué et al., 2015). Given two binary representations pertaining to two molecules, A and B, the Tanimoto coefficient calculates structural similarity as a function of the overlap of binary attributes set to '1', N, as illustrated in the equation below.

$$T(a,b) = \frac{N_{A\&B}}{N_A + N_B - N_{A\&B}}$$

The Tanimoto coefficient takes the range 0-1, 0 meaning no structural similarity and 1 meaning very structurally similar. It is important to note that a Tanimoto value of 1 does not necessarily mean that the two molecules being compared are the same. Enantiomeric molecules, such as Thalidomide, are an example where two molecules will have a Tanimoto coefficient of 1 although the two enantiomers differ chemically.

2D fingerprints are some of the most commonly used descriptors of molecular structure because they are fast to compute and are information-rich in terms of differentiating molecules according to bioactivity. Molecular fingerprints are binary representations of molecular structure which encode the presence or absence of chemical features in a molecule. 2D fingerprints can be divided into two main groups: predefined fragment dictionary based and those based on hashing methods (Leach & Gillet, 2007). In dictionary-based fingerprints, often referred to as structural keys, each bit represents a specific substructure from the predefined library. A common example is the MACCS keys fingerprint which contains 166 predefined substructures (Durant, Leland, Henry, & Nourse, 2002). While such fingerprints are widely used in substructure searches, one of their disadvantages is that they may not encode novel substructures.

Hashed fingerprints on the other hand are not based on a predefined fragment library which means that any substructure in any given molecule may be encoded (Leach & Gillet, 2007). However, these fingerprints are difficult to interpret because it is not possible to identify the specific substructure encoded by a specific bit in a binary string. Their main use is for calculating molecular similarity rather than performing substructure searching (Rogers & Hahn, 2010a). Examples include extended connectivity fingerprints (Rogers & Hahn, 2010b) which encode the neighbourhood of each atom within the molecule, and Daylight fingerprints (Butina, 1999) which encode atom pair information.

1.5.2 Cluster analysis

Molecular clustering is the process by which groups of molecules are put into clusters such that structurally similar molecules are found in the same cluster while dissimilar molecules are found in different clusters (Leach & Gillet, 2007). The first step in cluster analysis involves the generation of molecular descriptors. The second step involves using a metric for calculating the similarity or dissimilarity between all molecules. A clustering algorithm is then used in the third step to put molecules into clusters. Because 'distance' is used, this method is sometimes referred to as a distance-based measure approach of grouping molecules. When binary descriptors (fingerprints) are used, similarity coefficients (S) can be calculated and the distance is calculated as 1-S (Leach, 2001; Leach & Gillet, 2007).

Cluster analysis is a common cheminformatics technique that is usually applied to diversity selection (Khanna & Ranganathan, 2011). The similar property principle, as stated previously, is one reason why one would want a diverse set of molecules in a dataset of interest. If structural similarity corresponds to similar biological activity, a diverse set of molecules allows for wider coverage of activity space. In some cases cluster analysis is used to reduce the number of compounds of a certain chemotype within a dataset which is useful when one requires equal representation of chemotypes. This is the case in structure-activity relationship studies where overrepresentation of a particular chemotype can lead to bias in results.

Clustering methods can be divided into overlapping and non-overlapping methods, the most common of which are non-overlapping methods where each molecule belongs to a single cluster. With overlapping methods, a single molecule can belong to different clusters. Non-overlapping methods can also be divided into two groups: hierarchical and non-hierarchical; hierarchical clustering places molecules into clusters of increasing size for which dendrograms may be used to visualize (Leach, 2001). A common method of clustering, agglomerative hierarchical clustering, begins with each molecule in its own cluster and progresses by combining the most similar structures together until all molecules are placed in one huge cluster (Saeed, Salim, & Abdo, 2013)

There are different methods for calculating the similarity between clusters: single linkage, complete linkage or average linkage (Leach & Gillet, 2007). In single linkage methods, the distance between clusters is calculated as the minimum distance between any two compounds from each cluster. In complete linkage, the furthest distance between any two molecules represents the distance between the two clusters. The average distance between all pairs of molecules represents the distance between clusters in average linkage methods. The assignment of molecules to clusters often requires one to make a choice between specifying the number of clusters required (less than the number of molecules) and specifying a distance threshold (the distance under which molecules are assigned to the same cluster).

1.6. Quantitative structure activity relationships (QSAR)

A QSAR is a quantitative (mathematical) model that relates structure-derived properties of a molecule to its biological activity (Newby, 2014; Yee & Wei, 2012). QSAR modelling is based on the similar property principle which states that structurally similar molecules have similar biological activity. One advantage of QSAR models is that they allow for predictions to be made before a compound is synthesized and thus serve as a time and money saving tool in drug discovery (Park et al., 2014).

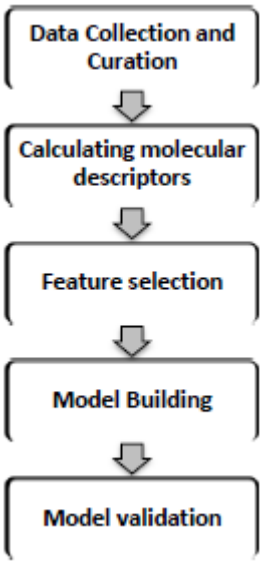

Figure 3: The process of creating QSAR models (Newby, 2014)

As illustrated in Figure 3, the process of developing QSAR models begins with collecting data for the property to be modelled. The quality of data is an important point for consideration as poor quality data often result in poor models. The next step is calculating molecular descriptors which describe the structural information of molecules. Various types of molecular descriptors exist and not all are relevant for a particular task. Therefore, feature selection is carried out to remove redundant and uninformative descriptors. For validation of the model, the dataset is often divided into training and test set prior to model development.

A wide range of QSAR methods for addressing different problems have been developed (Hou et al., 2006). QSAR models generally fall into 2 categories: regression models and classification models (Leach, 2001). Regression methods aim to give a real value output whereas classification methods aim to predict the class or group to which molecules belong (e.g. inhibitor or non-inhibitor) (Fujita, 1995). Multiple linear regression (MLR) is an example of a continuous model which deduces the linear relationship between biological activity (or any property of interest) and molecular descriptors. Problems may occur with MLR when the number of correlating descriptors increases, thus making efficient feature selection important. Artificial neural networks are an example of a classification method, the idea of which is inspired by the organisation and functioning of the central nervous system

(Hou et al., 2006). Other examples of classification QSARs include Logistic regression, Naïve Bayes, and the Random Forest classifiers (Dehmer et al., 2012).

1.6.1 Data collection and curation

The process by which a QSAR model is derived can greatly influence its utility as a predictive tool. Quality of data used can influence how models perform therefore it is important that data is of high quality. It is often the case that data used in QSAR model generation is obtained from different sources which makes the curation process more important. Curation is carried out to remove errors in the datasets. The standardisation of molecular structures is an important step in data curation. Incorrect structures can lead to calculation of wrong descriptors and this will subsequently render the models less effective if useful at all. The removal of duplicate molecules is also important as having duplicates in the model generation process can lead to overfitting and artificially high predictions which may not translate to external datasets.

1.6.2 Molecular descriptors

Mathematical representations of molecular structures, more commonly known as molecular descriptors, are used to deduce the relationship between structure and biological activity (Newby, 2014). A wide range of molecular descriptors exist ranging from experimentally obtained measurements such as solubility and logP to theoretically obtained values based on quantum and molecular mechanics descriptors. Binary representations of molecular structure more commonly known as molecular fingerprints can also be used as molecular descriptors in QSAR modelling. The choice of molecular descriptors used depends on the problem to be addressed. For example, logP (as a measure of lipophilicity) is often used as a descriptor to predict the bioactivity and membrane permeability of molecules.

1.6.3 Feature selection

The abundance of descriptors that can be used in QSAR modelling makes it necessary to have a method of selecting important descriptors for the problem at hand (Eklund, Norinder, Boyer, & Carlsson, 2014; Tuv, Borisov, Runger, & Torkkola, 2009). Not all descriptors carry necessary information for generating a QSAR model.

Three types of descriptors may be found in a descriptor pool: relevant, irrelevant and redundant descriptors (Anderson, Michalski, Carbonell, & Mitchell, 1983). Irrelevant descriptors are considered unimportant for the problem at hand while redundant descriptors are those that carry the same information e.g. highly correlated descriptors. Feature selection is the method by which irrelevant and redundant descriptors are removed from the descriptor pool to leave just the relevant ones for the problem at hand. Descriptor selection may be carried out by filter, embedded and wrapper methods. Filter methods rank the descriptors according to some metric and removes descriptors that are found to be irrelevant (Eklund et al., 2014). Examples of filter methods include the removal of highly correlated and low variance descriptors. Wrapper methods combine the learning algorithm with the feature selection by searching for subsets of descriptors from the original pool. The best subset is the one that produce the learner with the highest predictive accuracy (Tuv et al., 2009). Embedded feature selection methods look for the best subset of descriptors by embedding the feature selection process in model building (Eklund et al., 2014).

1.6.4 Model building

For generating a QSAR model, the intended use and purpose of the model have to be established. This is an important step that helps in choosing the best method often with a trade-off between predictive performance and interpretability (Newby, 2014). QSAR models are generally classified into unsupervised and supervised methods.

Unsupervised methods are so called because the learning process does not require one to distinguish between independent and dependent variables. This makes unsupervised methods ideal for discovering unknown patterns in a dataset, based on the similar property principle as mentioned previously (Dehmer et al., 2012; Peironcely, Reijmers, Coulier, Bender, & Hankemeier, 2011). Clustering is an example of an unsupervised technique often employed for selecting diverse subsets and for splitting datasets into training and test sets for supervised methods (Saeed et al., 2013).

Supervised methods require the dependent and independent variables to be distinguished and are more commonly applied in QSAR and ADMET modelling than unsupervised methods. Molecular descriptors make up the independent variables that are used to predict the dependent variable (biological activity or other property of interest). Supervised methods can be further divided into classification or regression methods. In classification methods, the predicted variable is categorical whereas regression methods predict numerical values (Dehmer et al., 2012; Newby, 2014). The choice between regression and classification methods depends on a few factors such as the intended use of the models and the availability of quality data. For example, one may be interested in knowing whether molecules are permeable or impermeable and in that case a classification method may be preferable. In another instance, one may be interested in predicting the biological activity of a molecule and in that case, regression methods may be optimal.

1.6.5 Model validation

After models have been generated with the training set, a validation step is required to assess the predictive performance of the model on the test set. A measure of accuracy is required to assess the performance of QSAR models.

For regression methods, the Pearson correlation coefficient, r^2, is often used to measure the linear relationship between experimental and predicted data. Measurements such as root mean square error (RMSE) and mean absolute error (MAE) are often used to measure the performance of models (Yee & Wei, 2012).

For classification methods, accuracy statistics may be used to measure model performance. These statistics are based on calculating the number of correctly classified compounds. The overall classification accuracy measures the number of correctly classified compounds. Statistics such as specificity and sensitivity measure the classification accuracy associated with different classes. The Cohen's Kappa statistic is often used to measure overall classification accuracy taking into account the probability of correct classification by chance (Ben-David, 2008).

1.7 QSAR models of cellular permeability

Predicting ADMET properties of molecules in compound libraries before synthesis is the goal of predictive modelling. Rule based approaches are often used as filters to screen out compounds that have small chances of being drugs. An example of a rule based approach is Lipinski's rule of 5 which states that an orally administered drug is less likely to be bioavailable if it has molecular weight greater than 500 Daltons, more than 5 hydrogen bond donors, more than 10 hydrogen bond acceptors, and a logP less than 5. Lipinski, however, states that this rule is not applicable to molecules that undergo carrier-mediated transport (Lipinski et al., 2012). Many successful drugs have been reported to violate different components of this rule, suggesting the rule should be used as a guideline. An example is Talaprevir, a hepatitis C drug developed by Vertex. This compound violates all components of the rule of 5 but is still active upon oral administration. The rule based approach is oftentimes too general therefore necessitating the development of more accurate predictive models.

There is a large number of published articles looking at predicting the permeability of compounds using QSAR models (L. Chen, Yao, Yang, & Yang, 2005; Guangli & Yiyu, 2006; T. Hou et al., 2006; Refsgaard, Jensen, Brockhoff, Guldbrandt, & Christensen, 2005). Permeability values obtained from *in-vitro* models of human cell permeability can be used to generate predictive QSAR models. One such example is Pham The et al. who took a classification QSAR approach to predict Caco-2 cell permeability using a dataset of 674 compounds from over 250 published articles (Pham The et al., 2011). They used DRAGON descriptors (Borota et al., 2011) and linear discriminant analysis QSAR (LDA-QSAR) to create 21 predictive models, the best of which was built using 9 DRAGON descriptors. The obvious limitation is that there is no attempt to clearly distinguish between passive diffusion and carrier-mediated transport in the descriptors used. All compounds are either assumed to have the same permeation mechanism (passive diffusion) or the descriptors are expected to describe both passive diffusion and active transport processes.

Another Caco-2 model, developed by Guangli & Yiyu (Guangli & Yiyu, 2006), made use of descriptors from an open source software, Chemistry Development Kit (CDK),

and the Support Vector Machine (SVM) learning method. The resultant model from this study suggests that the number of hydrogen bond donors and properties relating to molecular surface area are important determinants of membrane permeability. There was a strong correlation between experimental and predicted permeability values in the test set (r=0.85) suggesting that the combination of the aforementioned descriptors and SVM methods is effective in predicting Caco-2 permeability coefficients. However, the limitations to this study, like most QSAR studies, can be attributed to the methodology applied. A small data set was used, rendering the model applicable to a small range of compounds with shared structural similarities. They made no use of an external data set in the validation step, instead opting to divide their original data set into training and test set. This is a major limitation given that the molecules used were obtained from two published studies therefore likely to contain a relatively small range of chemotypes. One would therefore expect the domain of applicability for such models to be limited.

Refsgaard et al. also carried out in silico modelling of intestinal membrane permeability (Refsgaard et al., 2005). Permeability coefficients of 712 compounds were obtained from in vitro MDCK and Caco-2 studies undertaken in a single laboratory. In the MDCK assay, permeability was measured in the absorptive direction (A-B) and both absorptive and secretory transport were measured in the Caco-2 assays. Substrates of efflux transporters were identified and removed from the dataset, leaving only molecules that undergo passive diffusion. 9 descriptors were chosen to represent lipophilicity, flexibility, hydrogen bonding capability and steric properties. The best performing model is reported to have a 15% misclassification rate, with no compounds in the non-permeable group being misclassified. Despite the low misclassification rate, this study is a good example of how passive diffusion is the main focus of QSAR models of permeability and how, in some cases, substrates of membrane transporters are excluded from datasets.

Both passive diffusion and active transport processes are considered important to the permeability of compounds. Evidence from published articles suggests that most studies do not attempt to distinguish between carrier-mediated transport and passive diffusion in QSAR models. Because carrier-mediated transport is important

to the permeability of molecules, the accurate modelling of carrier-mediated transport is therefore an area which needs addressing.

1.8 Carrier mediated transport descriptors in modelling of cellular permeability

Most predictive modelling studies are focused on describing passive diffusion of molecules across cell membranes. This is unsurprising given that it is widely perceived that active transport has a small contribution to the ADME properties of most drugs compared with passive diffusion. There are also limitations which hinder active transport modelling such as availability of structural data and the variation of results between different labs. The promiscuity of transporters also hinders the process of creating models because of the many possible binding modes and the subsequent wide range of features that may render a molecule complementary to any given uptake or efflux transporter (Giacomini et al., 2010). Generalizable predictive models of active transport are therefore rare, with most studies limited to individual transporters (Cabrera, González, Fernández, Navarro, & Bermejo, 2006; Sedykh et al., 2013). The permeability glycoprotein (P-gp) is an efflux transporter that many studies have focused on (Bikadi et al., 2011; Broccatelli et al., 2011). This is most likely because many drug molecules are less efficacious if they are substrates of P-gp and therefore early identification of such substrates is important. However, most of the studies on single transporters make use of a 3D structure of the transporter in order to identify interaction points and create 3D QSAR models (Bikadi et al., 2011). QSAR modelling of individual transporters is an alternative for predicting compounds likely to interact with a particular transporter. However, developing models for all membrane transporters would be time consuming and near impossible given the lack of structural data.

With the use of various learning methods and descriptors from MOE (Chemical Computing Group) and DRAGON, Sedykh et al. (Sedykh et al., 2013) created QSAR models for use in predicting molecules likely to be transporter substrates. This is one of a few reports attempting to model active transport systems collectively instead of

focusing on individual transporters. The models generated show high external validation (71-100%; 5-fold external validation) and therefore such methodologies could be applied in generating descriptors of carrier-mediated transport.

Recently Dobson et al. (Dobson, Patel, & Kell, 2009) proposed the use of 'metabolite-likeness' as a descriptor of active transport. The authors suggest that drug uptake is predominantly an active uptake process and that passive diffusion plays a minimal role. This goes against the current belief that passive diffusion is dominant. They suggest that endogenous molecules permeate cells through membrane transporters and thus molecules that are structurally similar to metabolites will interact with and be transported by the same carrier-mediated transporter systems. If active transport is the main method by which molecules permeate the cell membrane and metabolite-likeness is a good descriptor of carrier-mediated transport capabilities, those metabolites that show little similarity to drugs could provide new avenues for drug discovery. Drugs can be designed to be similar to these metabolites and consequently bind to membrane transporters that current drugs do not interact with.

1.9 Aims & Objectives

The aim of this project is to develop classification models of molecule permeability that include novel descriptors of carrier mediated transport and to evaluate their predictive performance. To achieve this aim, a set of standard physicochemical and structural property descriptors will be extended by inclusion of two novel descriptors based on predicted molecular properties related to cell permeability. Molecular similarity to endogenous metabolites and affinity for membrane transporters predicted by PIDGIN will be used as carrier mediated transport descriptors. The performance of models generated by using an extended set of descriptors will be evaluated by comparison to models developed from commonly used descriptors such as lipophilicity, solubility and structural keys.

Specific objectives are:

- Assess the applicability of metabolite-likeness as a potential descriptor of likelihood of molecules to undergo carrier mediated transport by comparing structural similarity of approved drugs and commercially available compounds to endogenous metabolites. This objective is addressed in Chapter 3.1.
- Assess the applicability of a target prediction tool, PIDGIN, to predict potential substrates of membrane transporters. This objective is addressed in Chapter 3.2.
- Develop classification models of permeability that incorporate the novel descriptors and evaluate whether inclusion of novel descriptors improve the predictive performance of such models. This objective is addressed in Chapter 4.

2. MATERIALS AND METHODS

2.1 Software

A wide range of computational tools were applied to carry out different aspects of this study. The data mining tool, KoNstanz Information MinEr (KNIME) (Berthold et al., 2009; Warr, 2012), was used as the main platform for drug discovery and development applications. KNIME is a workflow environment that allows integration of machine learning, data mining and cheminformatics algorithms, adopting a pipeline concept where information flows between nodes connected by edges (Figure 4). Many software vendors such as ChemAxon (www.chemaxon.com) and RDKit (Riniker & Landrum, 2013; Saubern, Guha, & Baell, 2011) have implemented cheminformatics toolkits for use in KNIME. The ChemAxon toolkits were used in this study to convert SMILES strings into 2D structures, to carry out structure standardization and to calculate molecular descriptors and fingerprints. The RDKit toolkit also includes a fingerprint calculator with several different fingerprint types.

The Waikato Environment for Knowledge Analysis (WEKA) data mining package, developed at the University of Waikato, is a collection of machine learning algorithms that are useful for predictive modelling (Hall et al., 2009). Included in the WEKA collection are Logistic regression, Naïve Bayes and Random Forest classifiers for which KNIME nodes have been developed. These classifiers are used in this study and are described in detail in later sections. For data analysis and visualisation, Spotfire (www.spotfire.tibco.com) and Microsoft Excel were used.

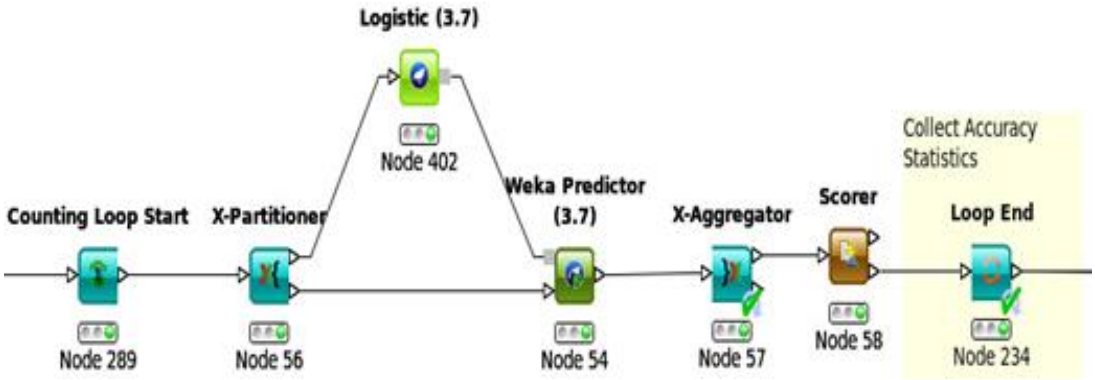

Figure 4: An example KNIME workflow illustrating the flow of information between nodes through edges

2.2 Dataset collection and curation

Simple molecular input line entry system (SMILES) strings of metabolites were obtained from the Human Metabolome Database (HMDB) v3.6 (Wishart et al., 2007). The data was obtained in an XML format from HMDB. In order to retrieve the required information, XSLT files were written to extract the HMDB ID, the SMILES string and information regarding the type of metabolite (e.g. endogenous, drug or food metabolite). The database contained a total of 41514 metabolite structures. Metabolites in this database were labelled according to the source from which they are derived. Included in the database are metabolites from plant sources (53), microbial metabolites (18), food metabolites (5819), drug metabolites (909), drugs (1509) and toxin/pollutant metabolites (43). 5653 metabolites were unlabelled.

SMILES strings of Food and Drug Administration (FDA) approved drugs were obtained from DrugBank (Wishart, 2006). The database contained 7759 molecules, both small molecule and biotech drugs, labelled as either withdrawn, nutraceutical, investigational, illicit, experimental, or approved. Of particular interest to this study were orally administered small molecule drugs. All other compounds with labels such as investigational-withdrawn, biotech, nutraceutical etc. were removed from the approved drugs list. The chemical structures of 100 000 commercially available compounds were obtained from the proprietary Evosource library (Evotec UK Ltd).

For the main purpose of this study, to evaluate whether the inclusion of carrier-mediated transport descriptors improves the predictive accuracy of permeability

predictions, a dataset containing 969 unique compounds with experimentally determined caco-2 apparent permeability (P_{app}) values in both apical to basolateral (A-B) and basolateral to apical (B-A) directions was provided by the DMPK group at Evotec.

It is often the case that experimentally obtained values are numerical. In classification QSAR methods, it is necessary to define thresholds in order to assign molecules into binary classes. Three classification schemes were applied for compounds in the caco-2 dataset. Caco-2 permeability cut-off values were applied in both A-B and B-A directions. Compounds were assigned to the "permeable" class in both directions if $P_{app} \geq 5 \times 10^{-6}$ cm/s whereas "non-permeable" was defined for compounds with $P_{app} \leq 1 \times 10^{-6}$ cm/s. Molecules with efflux ratio ($P_{app\ (B-A)}/P_{app\ (A-B)}$) ≥ 10 were classed as "high efflux" whereas compounds with efflux ratio ≤ 1 were classed as "low efflux."

In drug discovery, the objective is to develop QSAR models that are generalizable to many chemotypes. A generalizable model is one that is capable of predicting the relevant property of a novel molecule despite the chemotype. The overrepresentation of a particular chemotype or chemical series in a training set can lead to overfitting the model. This can have major influences on the overall performance of any model when it comes to predicting molecules from different chemotypes. We therefore hypothesize that having a representative set of compounds for each chemotype can give a better outlook at how the model is likely to perform when used to predict molecules from chemotypes that are unrelated to the training set chemotype. This model is likely to be more generalizable as it avoids over-fitting the model to a particular chemotype.

To analyse this, compounds in the caco-2 dataset were clustered and the centroid molecule from each cluster was used for subsequent QSAR model generation. The first step in the clustering process was to create a distance matrix from the fingerprints of each molecule using the Distance Matrix calculate node in KNIME. The commonly used extended connectivity fingerprint (ECFP) fingerprint encoding was used in this case. The ECFP takes into account the environment of each atom in a

molecule by encoding the neighbouring atoms of each individual atom up to a predefined diameter.

A distance matrix can be used for hierarchical clustering. The assignment of molecules to clusters requires the user to choose between a specific number of clusters or a distance threshold under which compounds are assigned to the same class. A threshold was chosen to ensure that a reasonable proportion of clusters contained compounds from the same class. The distance threshold selected was 0.3 and this means molecules that are within a distance of 0.3 from each other are assigned to the same class. Some clusters will contain more compounds than other clusters but the fact that a centroid molecule will be selected means that each cluster is equally represented.

Due to the majority of approved drugs being small molecules (Wishart et al., 2007), 1000 Daltons was set as the maximum molecular weight for any compound in the approved drug, metabolite and library compound dataset. A minimum molecular weight of 150 Daltons was also applied to the same datasets. This constraint was applied to remove ions and to ensure the inclusion of only small organic molecules containing at least 5 atoms. While this process reduces the chemical space covered, it ensures that the distribution of molecular weight across the datasets is the same. This is particularly important in this study as molecular fingerprints are employed for calculation of similarity between compounds from different datasets. The presence or absence of certain chemical substructures will be used to determine structural similarities. Fingerprint-based similarity measures are size dependant because large compounds are likely to have more substructures while small molecules have fewer substructures. In a molecular fingerprint, the presence or absence of a substructure in a molecule is indicated by setting the relevant bit to 1 or 0 (Holliday, Salim, Whittle, & Willett, 2003). Large molecular weight molecules are more likely to have more bits set to 1 because they contain more substructures and are therefore more likely to have artificially high similarity values to smaller molecules. Holliday et al. analysed the effects of molecular size on various similarity coefficients when used to measure the structural similarity between pairs of molecules and came to the conclusion that bit density affects similarity measures (Holliday et al., 2003).

Standardization of structures in all datasets was carried out using the Chemaxon Standardizer tool which is available as a KNIME node. The process of standardisation was carried out as suggested by Dobson et al.(Dobson et al., 2009) and involves aromatization, removal of salts, selection of the largest fragments and standardization of stereochemistry. The standardization of molecular structures is an important process that ensures that the same molecule can be recognised when represented by different chemical forms (Figure 5).

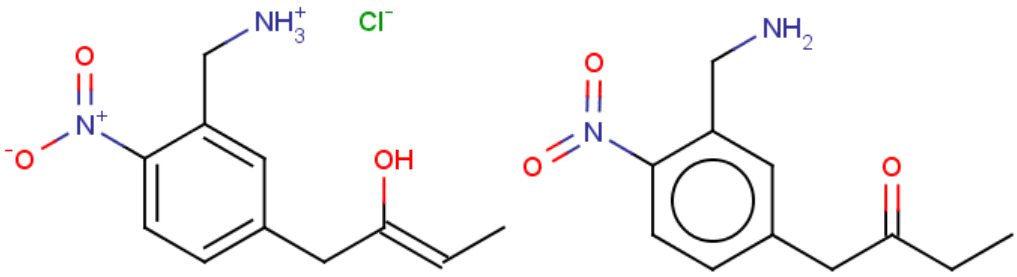

Figure 5: An illustration of how the same molecule can have different chemical representations.

As we are interested in the behaviour of molecules in biologically relevant environments, it is essential to use the biologically relevant representations. In all datasets, the most prevalent species at pH 7.4 was calculated and retained using the Major Microspecies at pH 7.4 node in KNIME. This represents the molecule that is most prevalent under biological conditions. The last step in the pre-processing step was the removal of duplicates. The GroupBy node was used for this purpose.

2.2.1 Summary of datasets

After the manual curation process the datasets contained 93 617 commercially available compounds (Evosource), 1349 small molecule FDA approved drugs and 24 113 HMDB endogenous metabolites. The caco-2 permeability dataset contained 969 compounds with experimentally determined P_{app} values in both A-B and B-A directions.

2.3 Molecular descriptors

2.3.1 Physicochemical and Structural Property Descriptors

In QSAR modelling, it is important to select descriptors that are relevant to the biological process under consideration. A series of Chemaxon descriptor calculators were used to generate molecular descriptors of the compounds in the caco-2 permeability dataset. Basic molecular values related to the elemental composition of the molecule were calculated by the Elemental Analysis node. For example, the molecular weight of each molecule was calculated using the Elemental Analysis plugin. Characteristic values related to the topological structure of the molecules were calculated using the Topology Analysis node. Hydrogen bond donor and acceptor counts were calculated using the H Bond Donor/Acceptor node. The pKa values were calculated using the pKa calculator plugin which calculates the pKa values of all atoms that can lose or gain a proton based on the distribution of partial charges. Molar Refractivity values were calculated using the molar refractivity node. LogD, logP, polar surface area (PSA) and molecular surface area were also calculated. Solubility coefficients of each molecule were calculated using MOE descriptors because of lack of a license for the relevant ChemAxon tool. MACCS keys counts were also generated for each of the compounds in the dataset. There are a large number of possible molecular descriptors that could have been calculated. One of the objectives of QSAR modelling is to ensure model interpretability. Therefore, the majority of descriptors that are difficult to interpret were not considered for predicting the permeability of compounds in a caco-2 cell.

To summarise, the descriptors calculated include size descriptors (molecular weight), lipophilicity (logP and logD) protonation descriptors (pKa), aqueous solubility (logS), geometrical descriptors (topology, molecular surface area & polar surface area), and structural descriptors (166 MACCS keys).

2.3.2 Novel Descriptors of Carrier-Mediated transport

In order to evaluate whether the inclusion of descriptors of carrier mediated transport can improve the performance of QSAR models of caco-2 permeability, there is a need for novel descriptors of carrier mediated transport. The two potential

sources of general descriptors of carrier-mediated transport pursued in this study are the concept of metabolite-likeness and the availability of a target prediction tool PIDGIN. The term 'potential sources' has been used on purpose to signal that these are not approved sources but they have the potential to serve as source descriptors of carrier-mediated transport and their utility for such will be investigated. Target prediction tools can be considered a good source of carrier-mediated transport descriptors if the tool has sufficient models of membrane transporters. The majority of studies are currently looking at predicting the likelihood of a molecule binding to a specific membrane transporter. For example, the permeability glycoprotein (P-gp) is a commonly known membrane transporter that is involved in the efflux of many compounds. For this reason, many seek models to predict the likelihood of molecules binding to P-gp in order to avoid those particular compounds as they are likely to be ineffective because they are removed from the cell. Here, we seek a general descriptor of carrier-mediated transport that can be utilised as an attribute or descriptor for the purposes of QSAR studies.

2.3.3 Metabolite-likeness as a descriptor of carrier-mediated transport

To assess the metabolite-likeness of approved drugs, various fingerprints of compounds from the endogenous metabolite and approved drug datasets were generated using the RDKit Fingerprint calculator node in KNIME. The choices of fingerprints available include connectivity fingerprints (Morgan and FeatMorgan), atom-pair fingerprints (Atom-Pair), topological torsion fingeprints (Torsion, RDKIT), avalon and substructure based fingerprints (Layered and MACCS keys). All fingerprint types were calculated and for each fingerprint, the Tanimoto similarity of each approved drug to each metabolite was calculated using a fingerprint similarity calculator node in KNIME. The Tanimoto coefficient is calculated as follows:

$$T(a,b) = \frac{N_{A\&B}}{N_A + N_B - N_{A\&B}}$$

where $T(a,b)$ represents the Tanimoto similarity between compounds a and b, $N_{A\&B}$ represents the number of on bits in common, N_A represents number of on bits in compound a and N_B represents the number of on bits in compound b. As a result, the Tanimoto similarity between any pair of compounds is always between 0 and 1,

1 meaning very similar. Each molecule from the drug set was compared to each metabolite (Figure 6) and the highest Tanimoto coefficient for each drug was retained as a possible descriptor termed the *nearest metabolite Tanimoto similarity (NMTS)* (Appendix Figure A. 1).

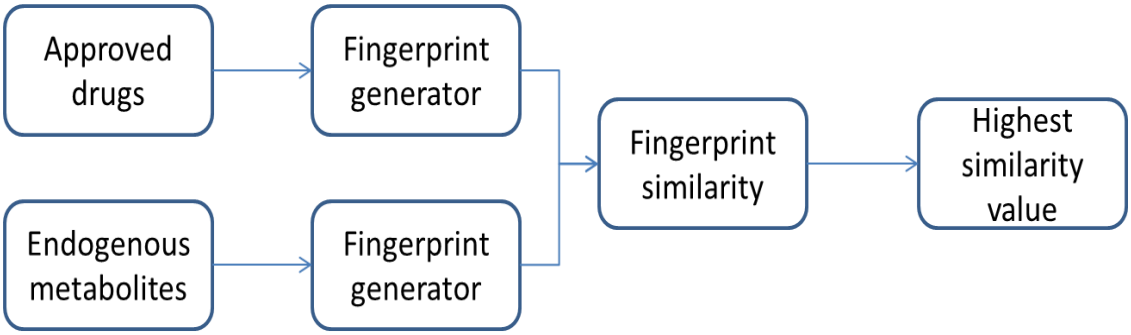

Figure 6: Illustration of fingerprint based metabolite similarity calculation

For the purpose of selecting the most appropriate fingerprint for this new descriptor, the metabolite-likeness of approved drugs and commercially available compounds was investigated with substructural (MACCS keys) and Chemaxon connectivity fingerprints (ECFP with diameter of 4 and Chemical Fingerprint). The commercially available compound dataset (Evosource) represents the general chemical space considered in drug discovery projects. If approved drugs owe their success to their metabolite-likeness, we hypothesize that the chemical space occupied by Evosource compounds should not be closer than the drug space to the metabolite space.

For each fingerprint encoding, the Tanimoto similarity of each approved drug and library compound to each endogenous metabolite was calculated as mentioned above (Appendix Figure A. 1). Because the number of library compounds (93617) far exceeds the number of approved drugs (1349), 1000 samples of 1400 randomly selected commercial compounds were taken to roughly match the number of approved drugs. The percentage of approved drugs and the average percentage of library compounds (from the 1000 samples) above given *NMTS* threshold values were calculated. We hypothesize that approved drugs are more similar to endogenous metabolites.

2.3.4 Target prediction as a source of descriptors of carrier-mediated transport

PIDGIN (Prediction IncluDinG INactivity) is a Bernoulli Naïve Bayes algorithm based target prediction too l(Mervin et al., 2015). The models are based on activity data obtained from ChEMBL(18) and inactivity data from PubChem. The models were trained and evaluated by a 5-fold cross validation, achieving an average recall and precision of 67.7% and 63.8% for active compounds and 99.6% and 99.7% for inactive compounds. Compared to models based on active data alone, inclusion of inactivity data is reported to produce models with better recall and precision. The PIDGIN tool was therefore used in this study to predict membrane transporter targets for compounds in the caco-2 dataset. The PIDGIN tool was made available as a node in KNIME by the computational chemistry group at Evotec.

For the purpose of this study, an indication of the abundance of membrane transporter models included in the PIDGIN tool was required. It is also important to know which transporters are expressed and function in caco-2 cells. A caco-2 gene expression profile was obtained from the Gene Expression Omnibus (GEO) database(D. Sun et al., 2002) in a format compatible with Microsoft Excel (csv). Gene expression data was presented as robust multi-array (RMA) values which are log2 quantile normalised. Quantile normalisation is a statistical method for making two distributions identical for ease of comparison. The dataset also contained gene codes. Because the RMA values can range between 4 and 13, it was necessary to define an RMA threshold value above which the relevant protein can be considered sufficiently expressed and functional in the cell.

In this study, the minimum RMA gene expression value was calculated as:

$$minimum\ expression\ value = median\ expression\ + 1$$

Because the expression values are $\log_2$ transformed, adding a value of one to the median expression value means the minimum expression threshold is double the median expression value.

After removal of genes below the minimum expression value, the next step was to perform a reference filter to investigate which of the expressed genes encode membrane transporters. Uniprot Accession numbers and gene codes of human

membrane transporters were obtained from UniProt (Magrane & Consortium, 2011). With reference to gene codes, the UniProt Accession numbers of sufficiently expressed human membrane transporters were obtained and matched against Accession numbers from the PIDGIN tool. All but membrane transporter prediction models were removed from the output. This output was in binary format, 0 indicating non-substrate and 1 indicating non-substrate of the relevant transporters. Molecules from the Caco-2 dataset compounds were used as input for the PIDGIN tool.

2.3.5 Descriptor normalization

After calculation of molecular descriptors, descriptor normalization was carried out. As descriptors have substantial differences in their ranges and values, it is essential to carry out a normalization procedure to ensure equal weighting of descriptors (Faulon & Bender, 2010). Examples of descriptor normalization methods include min-max normalization, Z-score normalization and normalization by decimal scaling (Patro & Sahu, 2015). For the purposes of this study, min-max normalization was carried out to perform a linear transformation of all values such that the minimum and maximum values of each descriptor range between 0 and 1. The formula for a min-max normalisation is:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Where x' is the rescaled value, x is the original value, max (x) is the original maximum value and min (x) is the original minimum value (Patro & Sahu, 2015).

2.3.6 Feature selection

The initial descriptor pool contained over 200 descriptors. Different approaches to reduce dimensionality and redundancy in the descriptor set were taken for each QSAR classifier. Molecular descriptors with similar values between molecules (low variance descriptors) do not carry useful information. The variance of each descriptor was calculated and descriptors with variance below 0.1 were removed from the dataset. As variance is range dependant, it is important to carry out descriptor normalization before applying a variance filter.

After removing descriptors with low variance, highly correlated descriptors were removed from the dataset. A correlation filter was used to remove descriptors carrying the same information. Here we calculate the correlation coefficient between descriptors as Pearson's Product Moment Coefficient (r) and set a threshold of 0.85. For Naïve Bayes and Random Forest algorithms, low variance and correlation filter methods were applied. The correlation filter of 0.85 retained 144 descriptors including MACCS keys, physicochemical and the novel descriptors of carrier mediated transport.

Backward feature selection was used to select important descriptors, from the 144 descriptors, for the Logistic regression classifier (Appendix Figure A. 2). The backward feature selection for Logistic regression was performed three times to select important descriptors for A-B permeability classification, B-A permeability classification and efflux ratio classification. In this method, the classification algorithm is trained on the initial n (144) input features (descriptors), then removes one input feature at a time and trains the same model on n-1 input features n times. The input feature whose removal has produced the smallest increase in the error rate is removed, leaving n-1 input features. The process is then repeated using n-2 input features, and so on. Each iteration k produces a model trained on n-k input features and an error rate e(k). Selecting the maximum tolerable error rate, we define the smallest number of features necessary to reach that classification performance with Logistic regression. For the purpose of this study, the smallest subset of descriptors that produce an error rate below 0.1 was chosen.

2.4 Development of QSAR models of Caco-2 permeability

2.4.1 Classification Accuracy and Cohen's Kappa calculation

Overall classification accuracy can be defined as the proportion of correctly classified instances. It is calculated as:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

Where TP is number of true positives, TN is number of true negatives, FP is the number of false positives and FN is the number of false negatives. A positive

classification is user defined and in this case, Permeable and low efflux are defined as positives whereas non-permeable and high-efflux are defined as negatives. This means that a true positive or true negative classification corresponds to a compound that is correctly classified. A false positive is a misclassification where a negative is classified as positive and a false negative classification occurs when a positive is classified as negative. The possible outcomes of a classification problem can be described using a confusion matrix (Figure 7).

		Observed class: 1	Observed class: 0
Predicted class	1	True Positive (TP)	False Positive (FP)
	0	False Negative (FN)	True Negative (TN)

Figure 7: Confusion matrix illustrating the possible outcomes of a classification problem

Cohen's kappa is a statistic used in this study to measure the level of agreement between the actual class and the predicted class of compounds taking into account the likelihood of random agreement (Cohen, 1968). Cohen's kappa is calculated as:

$$K = \frac{P_0 - P_e}{1 - P_e}$$

Where P_0 is the observed agreement between actual and predicted class and P_e is the probability of random agreement between the actual class and the predicted class.

2.4.2 Training and validation sets

Instead of using a single split to create a training and validation set, the 5-fold cross validation technique was employed in the model building and validation procedure to evaluate the performance of models. This technique therefore requires the training of multiple models (5 models in a 5-fold cross validation). The dataset is split into 5 subsets of equal size and the following and each subset is used as the

validation set once and as part of the training set four times. This means each compound in the dataset is predicted once and used to train the model four times.

2.4.3 Convergence calculation

1000 iterations of the 5-fold cross validation technique was repeated using Logistic regression classifier. A cumulative moving average plot of A-B permeability classification accuracy was generated. The number of iterations after which the moving average accuracy converges was employed as the number of iterations to be performed for any subsequent models generated in this study. 20 iterations of the 5-fold cross validation procedure were performed for each QSAR model generation and validation.

For the purposes of this study, Logistic regression, Naïve Bayes and Random Forest classifiers were used. These methods are briefly described below. For each of these three classifiers, 3 datasets were generated. These datasets are the same apart from the set of descriptors used. In the standard descriptor dataset, descriptors that are considered relevant for studying drug cellular permeability were included. The second dataset consists of these standard descriptors and nearest metabolite Tanimoto similarity (NMTS) of each of the Caco-2 dataset compounds. The third dataset consists of standard and PIDGIN descriptors. Compounds were labelled as either permeable or non-permeable in both Apical to Basolateral and Basolateral to Apical directions. Furthermore, compounds were labelled as being high or low efflux according to the classification criteria previously described.

2.4.4 Logistic Regression Classifier

Logistic regression employs the concept of regression to perform binary classifications (Yee & Wei, 2012). The nature of the logistic function ensures that values obtained are between 0 and 1. The output can be transformed into a binary response based on a threshold value, in this case 0.5. This means any output greater than 0.5 is transformed to 1 and below 0.5 is transformed to 0.

The regression model is then generated as:

$$Logit\ (P) = b_0 + b_1x_1 + b_2x_2 \ldots + \ b_kx_k$$

Where Logit (P) is the logit transformation of the probability of one of the outcomes, $b_0, b_1 \dots b_k$ are regression coefficients and $x_1, x_2 \dots x_k$ are molecular descriptors. The Logistic regression classifier was implemented in KNIME as illustrated in Figure 8.

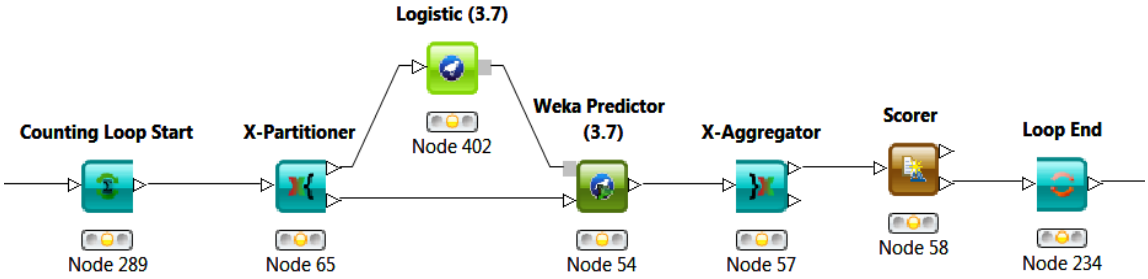

Figure 8: Logistic regression classification model development in KNIME

2.4.5 Naïve Bayes Classifier

Naïve Bayesian classifiers are based on the Bayesian theory of probability. Given a dataset with instances to be classified or in this case molecules to be classified according to their permeability or efflux ratio, Naïve Bayesian classifiers takes into account the prior probability of the molecule belonging to each of the classes. From this, the posterior probability can be calculated. The compound belongs to the class for which its posterior probability is greater than 0.5 (Yee & Wei, 2012). The equation for the Naïve Bayes classifier implemented in the WEKA package is as follows:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

Where P(A) is the prior probability of A (in this case it is the prior probability of a molecule being either permeable or low efflux), P(B|A) is the probability of B given A and P(B) is the probability of B (the descriptors). The Naïve Bayes classifier was used in this study as illustrated in Figure 9.

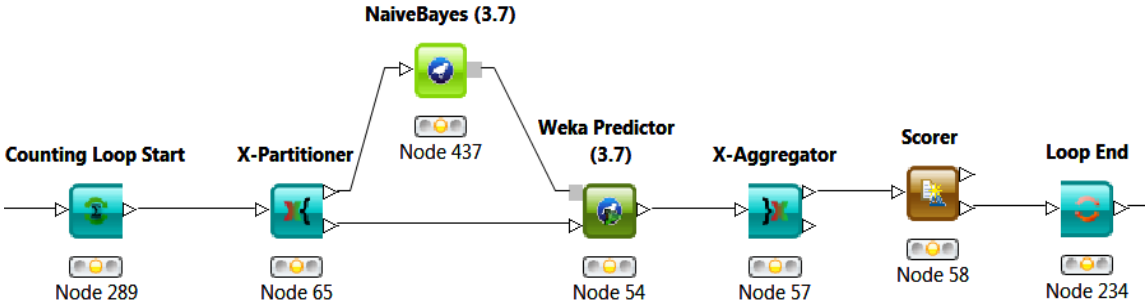

Figure 9: Naïve Bayes classification model development in KNIME

2.4.6 Random Forest Classifier

A Random forest classifier is based on an ensemble of decision trees. One advantage of Random forest classifiers is their inherent ability to handle high dimensional data (Svetnik et al., 2003). However, this has been reported to come at the cost of classification accuracy. In a 5 fold cross-validation, for each fold, a random forest of 10 trees is constructed. In the WEKA implementation, each tree is constructed while considering only 8 random descriptors from the available set of descriptors. The relevant descriptors of the test set compounds are passed through each tree and probabilities are calculated. Each compound is predicted to be in a particular class and if the probability is greater than 0.5, it belongs to that particular class. If the probability is lower than 0.5, it belongs to the other class. This means that the molecule belongs to the class for which its probability is greater than 0.5. As 10 trees are created, the compound's class is predicted 10 times. This means that the compound belongs to the class it is predicted to belong the most. For example, if a compound is predicted to be permeable by 6 trees and non-permeable by 4 trees, it belongs to the permeable class of compounds. The Random Forest classifier was implemented in this study in KNIME as illustrated in Figure 10.

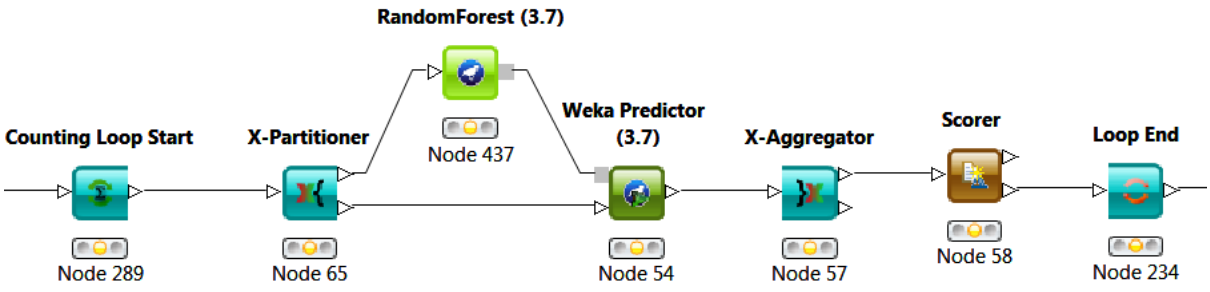

Figure 10: Random Forest classification model development and validation in KNIME

2.5 Statistical analysis of models and descriptors

2.5.1 Comparison of model performance

In order to evaluate whether active transport descriptors add value in terms of predictive accuracy and Cohen's kappa, independent samples t-tests were performed to compare the average accuracy and kappa of models generated from standard descriptors and the novel descriptors. The Independent samples t-test generates significance (p) values which are used in this study to evaluate significant differences between mean accuracy and Cohen's Kappa values. Figure 11 shows a schema of how this method was implemented in this study.

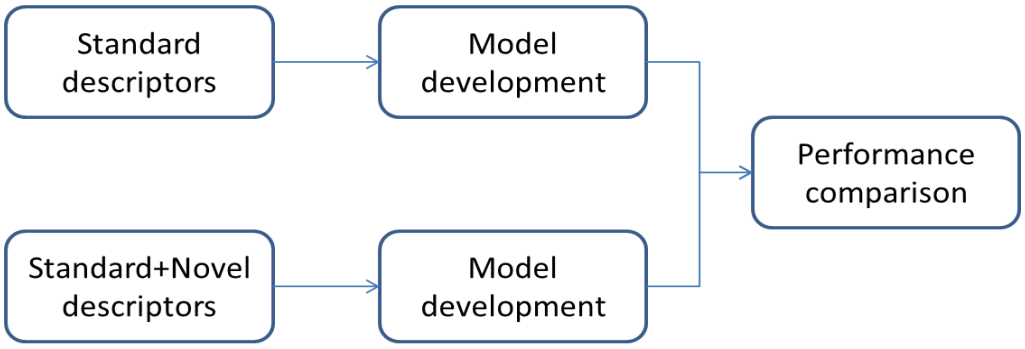

Figure 11: Schema for generating and comparing performance of models

2.5.2 Descriptor randomization

Each descriptor was randomized in turn in order to identify the descriptors that contribute the most to model performance. The classification accuracy of models when none of the descriptors are randomised is calculated. The non-randomised mean classification accuracy of models is then compared to the mean accuracy when each individual descriptor is randomised. An independent samples t-test was carried out to compare the means. Because of the number of comparisons to be made (equal to the number of descriptors), the problem of multiple comparisons arises. The Bonferroni's correction for multiple comparisons was therefore applied (Rafter, Abell, & Braselton, 2002).

Other methods such as simply leaving one descriptor out could be used but this can have an effect on the model performance simply because of an unequal number of descriptors. Another approach is to replace each descriptor with a random variable. This approach was not taken as the distribution of values cannot be maintained. The problem of multiple comparisons has to be addressed. Instead of using the significance (p) value of 0.05, the Bonferroni's critical value can be used. The Bonferroni critical value was calculated as α/n where α is the significance level applied for the independent samples t-test and n is the number of comparisons to be made (equal to the number of descriptors).

3. NOVEL DESCRIPTORS OF CARRIER MEDIATED TRANSPORT

3.1 Analysis of Metabolite-likeness

3.1.2 Results and Discussion

It has been proposed that structural similarity to endogenous metabolites can be used to measure the likelihood of molecules undergoing carrier mediated transport (Dobson et al., 2009). Endogenous metabolites are often substrates of membrane transporters. It was therefore hypothesized that approved drugs are successful because they can permeate the cell via membrane transporters. The structural similarity of approved drugs to endogenous metabolites, using various fingerprints available in the RDKit fingerprint calculator, was assessed and results are shown in Figure 12.

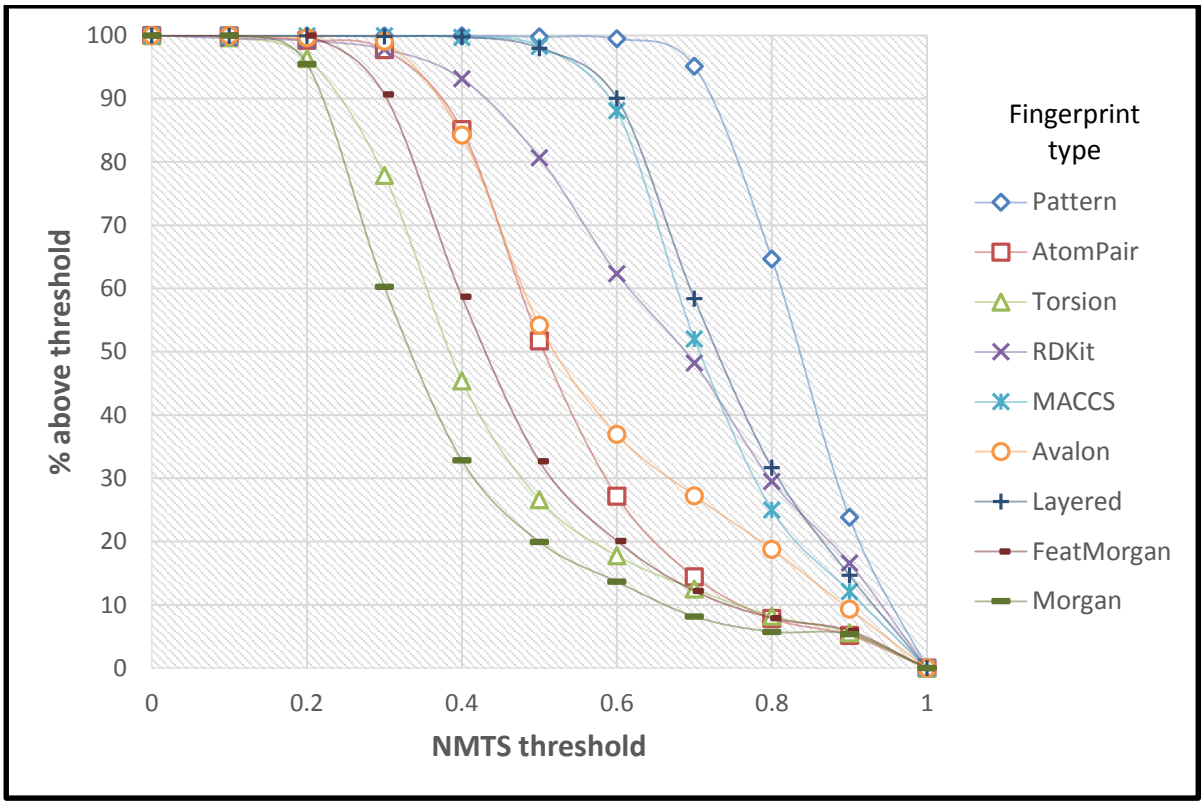

Figure 12: Percentage of approved drugs above a given NMTS threshold

The different fingerprint encodings produce different percentages of drugs above a given nearest metabolite threshold (NMTS). Drugs tend to show higher structural

similarity to metabolites when structures are encoded by structural keys fingerprints such as the Pattern, Layered and MACCS keys fingerprints. On the other hand, drugs appear less similar to metabolites when connectivity fingerprints such as Morgan, Torsion and FeatMorgan fingerprints are used. Around 98% of approved drugs are shown to have a nearest metabolite Tanimoto similarity greater than 0.5 when encoded by MACCS keys fingerprints (Figure 12). This is a larger percentage compared to the 90% observed by O'Hagan et al. (O'Hagan, Swainston, Handl, & Kell, 2014) and the difference is most likely because a different set of metabolites was used in this study. Considering that 24113 endogenous metabolites were used in this study and 1113 in the study by O'Hagan et al., it is unsurprising that the results of this study suggest a greater similarity of approved drugs to endogenous metabolites. Despite the differences between the studies, the same conclusions can be made: approved drugs are structurally similar to endogenous metabolites and the level of similarity depends on the fingerprint encoding used.

The similarity is much greater with structural keys compared with other types of fingerprints. This means that drugs are more similar to endogenous metabolites in terms of the substructures they contain but less similar when connectivity is taken into account. The findings of this study are in agreement with the idea that, for some molecules, the process of drug development is such that they are shifted towards the chemical space occupied by metabolites(Khanna & Ranganathan, 2011; Peironcely et al., 2011). O'Hagan et al. suggest that this process renders the drug more likely to interact with membrane transporters that transport the relevant endogenous metabolites. Metabolite-likeness could therefore be a useful descriptor of carrier-mediated transport for the purposes of QSAR studies of membrane permeability. However, it is not enough to say that approved drugs are similar to endogenous metabolites as chemical similarity is context dependent (Mestres & Maggiora, 2006).

Here, only the chemical space occupied by approved drugs is considered. It is also necessary to assess the metabolite likeness of molecules that span the general chemical space considered in drug discovery projects. Commercially available molecules are one such source. If metabolite likeness is a criterion to be used in the

enrichment of drug discovery libraries and if the drug development processes resembles a shift in chemical structure to that occupied by metabolites, we should be able to see significant differences in the metabolite likeness of approved drugs and commercially available compounds.

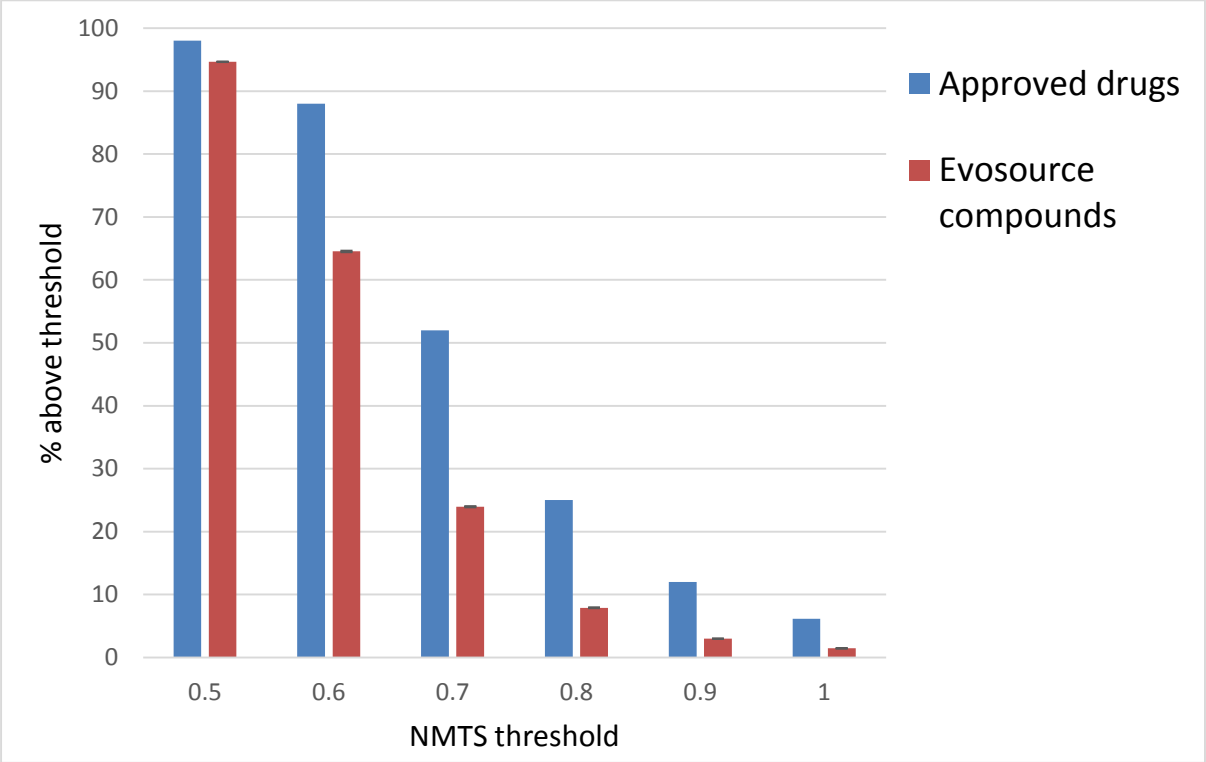

Figure 13: Comparison of percentage of approved drugs and library compounds that are above the specified NMTS threshold in the MACCS keys descriptor space. Bars representing percentages of Evosource compounds have standard error bars attached. These may not be visible because the standard error values are miniscule.

At all given NMTS thresholds (0.5-1.0), approved drugs are significantly more similar to endogenous metabolites than library compounds when the MACCS keys fingerprint encoding is used (Figure 13). 98% of approved drugs have a Tanimoto similarity greater than 0.5 compared with 95% of library compounds. Of the approved drugs, 88% have a NMTS value greater than 0.6 in contrast to the 65% of Evosource compounds. The reduction in percentages is greater for Evosource compounds than approved drugs. This suggests that there are about 12% of

approved drugs that are less similar to endogenous metabolites than 65% of library compounds.

Considering that 95% of commercially available compounds have an *NMTS* value above the 0.5 threshold suggested by O'Hagan et al. (O'Hagan et al., 2014) (Figure 13), the rule of 0.5 does not appear to be a useful criterion in distinguishing successful drugs from molecules from the general chemical space. However O'Hagan et al. did acknowledge that molecules that obey the rule will not necessarily be successful drugs (O'Hagan et al., 2014). The conclusion was that those that do not obey the rule are unlikely to be successful. This could be seen as misleading because the majority of compounds in the non-drug or library compound space also obey the rule. At higher NMTS values however, a greater separation between approved drugs and library compounds can be seen. This suggests that not all but a high proportion of approved drugs are more similar to endogenous metabolites than library compounds in the MACCS keys descriptor space.

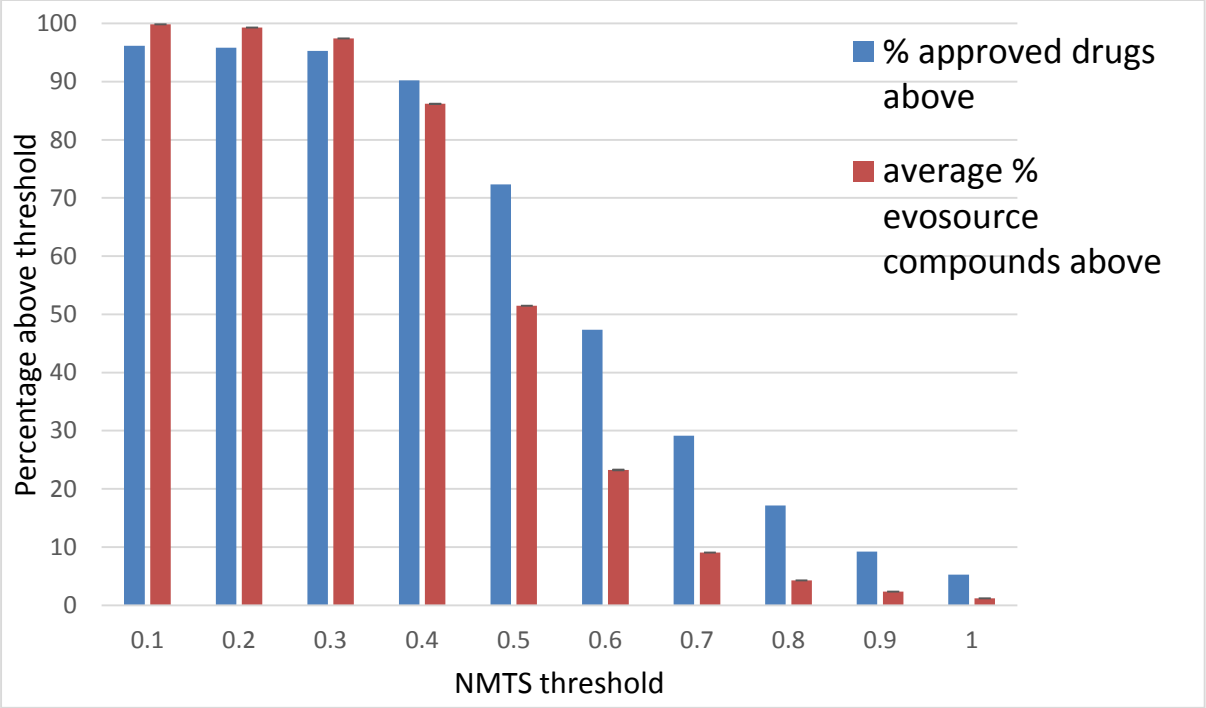

Figure 14: Comparison of percentage of approved drugs and library compounds that are above the specified threshold with Chemical Fingerprint encoding. Bars representing percentages of Evosource compounds have standard error bars attached. These may not be visible because the standard error values are miniscule.

A similar trend to that observed when MACCS keys fingerprints are used is also observed with the path based Chemaxon fingerprint. Between thresholds 0.1–0.3, the percentage of Evosource compounds is greater than that of approved drugs (Figure 14). This suggests that there are a few approved drugs that do not resemble endogenous metabolites. This is probably due to the inclusion of non-bioactive molecules in the list of approved drugs e.g. the diagnostic agent perflutren. However, these similarity values are too low for conclusions to be drawn. At higher NMTS thresholds (> 0.5), the differences become apparent. For example, more than 70% of drugs have a NMTS greater than 0.5 compared with about 50% of Evosource compounds. The overall trend shows that a higher number of approved drugs have greater similarity to endogenous metabolites compared with library compounds.

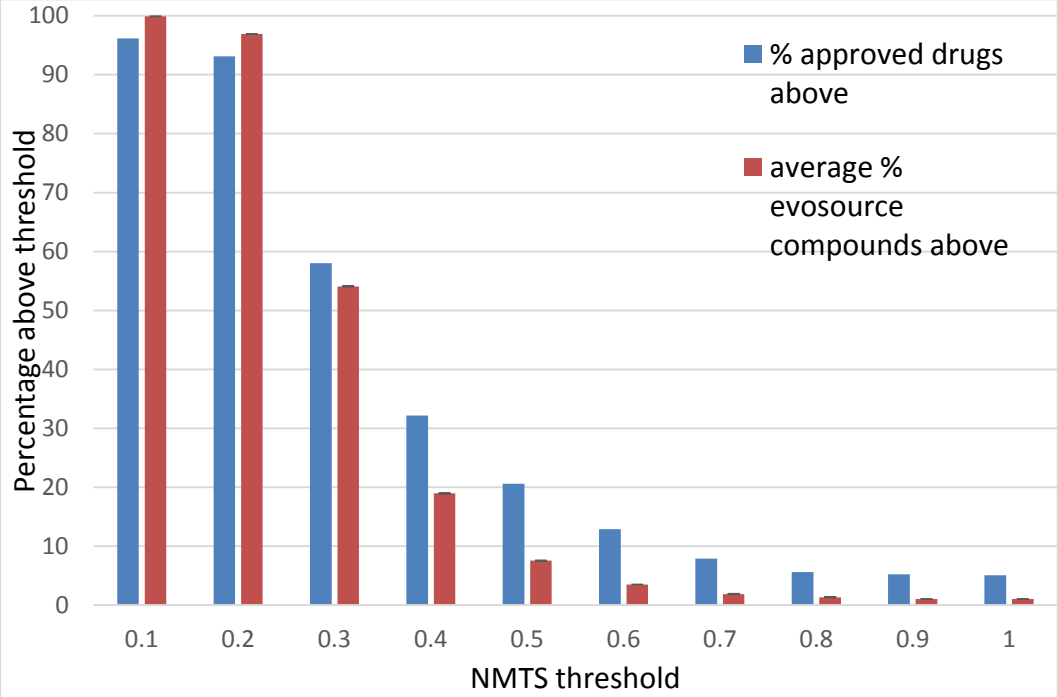

Figure 15: Comparison of percentage of approved drugs and library compounds that are above the specified NMTS threshold with Extended Connectivity Fingerprint encoding. Bars representing percentages of Evosource compounds have standard error bars attached. These may not be visible because the standard error values are miniscule.

When the circular connectivity (ECFP_4) fingerprint is used, approved drugs still appear more similar to endogenous metabolites compared with Evosource

compounds (Figure 15). However, the level of similarity is greatly reduced with the fingerprint encoding. 98% of approved drugs have NMTS greater than 0.5 with MACCS keys, around 70% with ChemicalFP, and 20% with ECFP_4. The most likely reason is that ECFP_4 encodes more chemical information by representing the presence or absence of particular substructure and their connectivity (Rogers & Hahn, 2010a).

3.1.3 Conclusions of Metabolite likeness

The nearest metabolite similarity values of both approved drugs and commercially available compounds differ with each of the three fingerprint encoding methods. However, approved drugs are consistently more similar to endogenous metabolites than are library compounds despite the fingerprint encoding. This suggests that approved drugs are generally more similar to endogenous metabolites than compounds spanning the general chemical space considered in drug discovery. Since molecular structure can be captured in several ways (Dobson et al., 2009), the three methods used to represent molecular structure and calculate molecular similarity show that metabolite-likeness is not just relevant to a particular structural encoding but is applicable to all the chemical spaces considered. This finding is in agreement with that of Dobson et al. (Dobson et al., 2009) who came to the same conclusion that approved drugs are more similar to endogenous metabolites than compounds found in typical screening libraries, despite the descriptor space considered. Because endogenous metabolites are often substrates of membrane transporters, it is reasonable to assume that molecules that are similar to endogenous metabolites are likely to undergo carrier-mediated transport. It was therefore hypothesize that metabolite-likeness can be used as a descriptor of a molecule's likelihood to undergo carrier-mediated transport. *NMTS* is therefore used in this study as a novel descriptor of carrier mediated transport.

3.2 Assessing applicability of PIDGIN, a Target Prediction Tool

The PIDGIN (Mervin et al., 2015) target prediction tool (described in detail in the methods section) was used to predict substrates of membrane transporters in the caco-2 dataset. As mentioned in the methods section, it was important to investigate which membrane transporters are expressed in caco-2 cells for which models are available in the tool. This section aims to give details of that investigation.

The majority of studies of this nature are differential expression studies aiming to identify differentially expressed genes between cells of interest e.g. normal and diseased cells (Hayeshi et al., 2008; Siissalo et al., 2007). Not many studies have pursued a minimum expression threshold and the reasons for this are wide ranging. One reason is that gene expression is a dynamic process which makes the process of choosing a minimum expression threshold difficult. Another difficulty stems from the fact that gene expression does not always correlate with protein levels.

However, a publication was found that carried out a similar study and came up with the idea of using the median of all expression values as the minimum detection threshold (Jin & Wang, 2009).

3.2.2 Results and Discussion

A minimum expression threshold robust multi-array (RMA) value of 6.5 was applied. Above this threshold, membrane transporters are considered sufficiently expressed and therefore more likely that the membrane transporters encoded function in Caco-2 cells. Below this threshold, we assume membrane transporters are not sufficiently expressed and therefore exhibit no function in caco-2 cells.

Table 1: Membrane transporters expressed in caco-2 cells with models available in PIDGIN

Uniprot Accession	Protein Names	Caco-2 RMA expression value
P49281	Natural resistance-associated macrophage protein 2 (NRAMP 2)/(Solute carrier family 11 member 2)	9.16
P53985	Monocarboxylate transporter 1 (MCT 1) (Solute carrier family 16 member 1)	7.62
P19634	Sodium/hydrogen exchanger 1 (APNH) / (Solute carrier family 9 member 1)	7.01
P31645	Sodium-dependent serotonin transporter (5HT transporter) (5HTT) (Solute carrier family 6 member 4)	6.81
Q9UNQ0	ATP-binding cassette sub-family G member 2 (Breast cancer resistance protein)	6.68
P33527	Multidrug resistance-associated protein 1 (ATP-binding cassette sub-family C member 1)	6.64

Only 6 PIDGIN models appear to be membrane transporter models that are expressed in the caco-2 cells according to the minimum expression RMA threshold of 6.5 (Table 1). It must be acknowledged that this is not an exhaustive list of membrane transporters and probably does not resemble the abundance of membrane transporters in caco-2 cells. Some of these are ion and metal transporters which may not be relevant to the transport of drug like molecules e.g. Natural resistance-associated macrophage protein 2 (NRAMP 2) and sodium/hydrogen transporter (APNH)(Cingolani & Ennis, 2007; Nevo & Nelson, 2006).

The Monocarboxylate transporter 1 (MCT1) is reported to transport short chain monocarboxylates and fatty acids. This means the protein is likely to be involved in the transport of small molecules that are of relevance to drug discovery. A study by Okamura et al. (2001) concerned with the uptake of nataglinide, an oral hypoglycaemic agent, in caco-2 cells suggests that MCT1 is indeed expressed and functions in caco-2 cells(Okamura, Emoto, Koyabu, Ohtani, & Sawada, 2002). This study also suggests that MCT1 is relevant to the transport of small molecules that are pursued in drug discovery projects.

The serotonin transporter (SERT) is widely reported to be expressed in caco-2 cells. Indeed caco-2 cells have been used to study the function of SERT(Iceta, Aramayona,

Mesonero, & Alcalde, 2008). Whether this transporter can mediate the transport of small molecule drugs is not clear. However, SERT has been reported to transport other molecules such as dopamine(Larsen et al., 2011). This suggests that SERT could indeed be responsible for transport of monoamines, as both serotonin and dopamine are monoamines.

The breast cancer resistance protein (BCRP) is an efflux transporter known to transport a diverse range of molecules. The protein is reported to be expressed in caco-2 cells and indeed caco-2 cells are used in studies of the mechanistic function of BCRP. Gene knockout studies in caco-2 cells are also carried out which can help to identify and avoid potential substrates of efflux transporters.

The multidrug resistance-associated protein 1 is part of a group of ABC proteins responsible for the efflux of chemically diverse group of molecules(Leslie, Deeley, & Cole, 2005). Evidence suggests that the MRP1 is expressed in caco-2 cells and therefore relevant to studies regarding carrier mediated transport in such cells (Prime-Chapman, 2004).

3.2.3 Conclusions

The caco-2 cell is reported to express a wide range of membrane transporters, a characteristic that gives the cell line resemblance to intestinal epithelial cells. However, an overview of the literature suggests that not many studies have tried to select a minimum expression threshold for caco-2 cells. The subsequent use of results of this study should therefore be interpreted with knowledge of potential limitations. The PIDGIN tool was therefore used in this study to predict substrates of the six membrane transporters listed above within the caco-2 dataset. Two of the membrane transporters listed are ion or metal transporters which may not be of direct relevance to permeability of small molecule drugs. The remaining transporters have been widely implicated in drug transport of a diverse range of molecules which suggests they are of relevance to membrane transport.

The models for each protein in the target prediction tool, PIDGIN, were developed using active and inactive compounds from ChEMBL and PubChem respectively (Mervin et al., 2015). Molecules that are active against membrane transporters, or

any protein for that matter, can be classified as substrates or inhibitors. The target prediction tool cannot distinguish between substrates and inhibitors of membrane transporters. This presents a potential limitation of using the target prediction tool to identify molecules that may undergo carrier mediated transport. These predictions were used in a binary fingerprint format as descriptors of carrier-mediated transport for the purposes of developing predictive QSAR models of caco-2 permeability.

4. DEVELOPMENT OF CLASSIFICATION MODELS

4.1 Caco-2 Permeability dataset

The aim of this section is to give an outline of the datasets used to generate QSAR models. The classification criterion mentioned in the methods section (section 5.2)has led to the selection of compounds with relevant permeability (Table 2).

Table 2: Number of compounds in Caco-2 permeability dataset

Direction	**Permeable count (P_{app} ≥ 1x10^{-6} cm/s)**	**Impermeable count (P_{app} ≤ 1x10^{-6} cm/s)**	**Unclassified**	**Total**
A-B	486	241	242	969
B-A	779	50	140	969

A

C

C

ording to the classification criteria mentioned in the methods section, 242 compounds were unclassified in the A-B direction whereas 140 compounds were unclassified in the B-A direction. This is because the P_{app} values of these compounds lie between 1x10^{-6} and 5x10^{-6} cm/s. A plot of A-B and B-A permeability suggests, in general, that there is no correlation between A-B and B-A permeability (Figure 16).

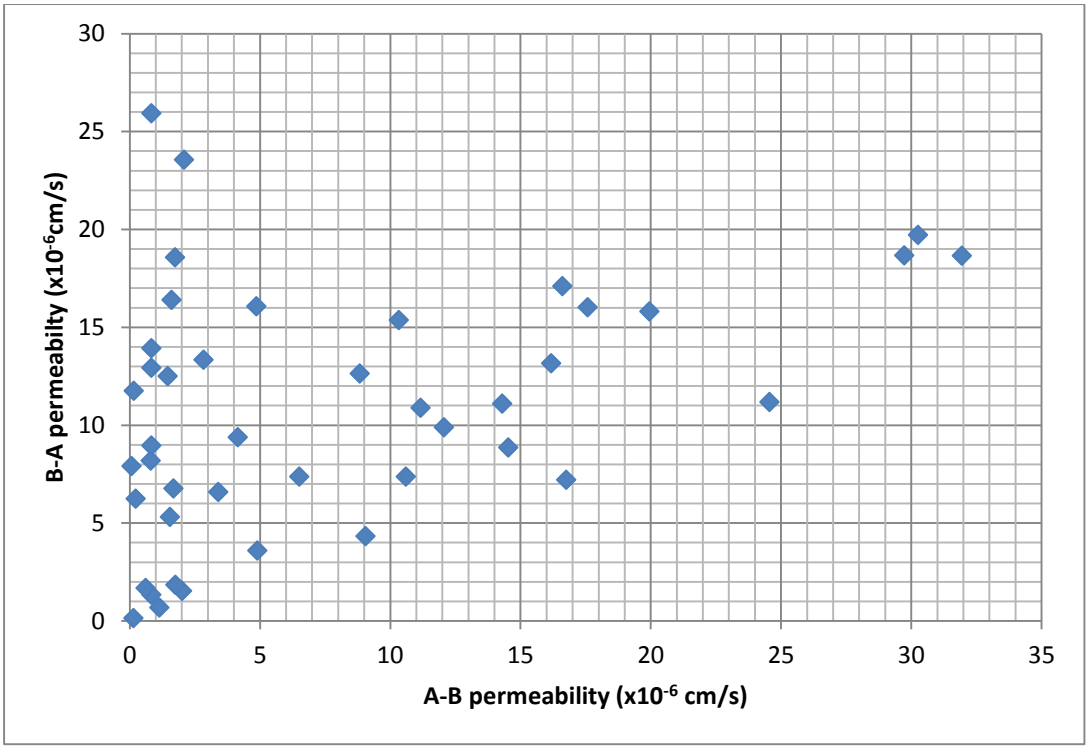

Figure 16: Plot of A-B vs B-A P_{app} values of caco-2 dataset

This shows that for this set of compounds, there are factors influencing the permeability of some compounds in the B-A direction that may not be relevant in the A-B direction. The most likely reason why more compounds are permeable in the B-A direction is because they interact with and are transported by efflux transporters. However, this could also be the reason why molecules with high B-A permeability values may have low A-B values.

A study of transport of epicatechin, a tea flavonoid with preventative properties for cancer, in caco-2 cells demonstrates how molecules can be more permeable in the B-A direction compared with the A-B direction (Vaidyanathan & Walle, 2001). Epicatechin was found to be impermeable in the A-B direction but slightly permeable in the B-A direction (P_{app} = 0.67 x 10^{-6} cm/s). According to the criteria used in this study, this molecule would be classed as impermeable in both directions. Interestingly, in the presence of an inhibitor of multiple resistance protein 2 (MRP2), an efflux transporter, permeability in the A-B direction is clearly measured (P_{app} = 0.31 x 10^{-6} cm/s). This suggests the impact of membrane transporters on both A-B and B-A permeability of molecules.

From the information available, the conclusion is that both passive diffusion and carrier mediated transport may be important for the permeability of molecules and in that case, QSAR models that incorporate descriptors of carrier mediated transport are expected to show better predictive performance to those developed from physicochemical and structural property descriptors alone.

Table 3: Number of compounds in the efflux class

High efflux	Low efflux	Unclassified	Total
135	367	467	969

According to the classification criteria (section 5.2), there were 135 high efflux and 367 low efflux compounds, making up a total of 502 compounds that have an efflux ratio that is either below 1 or above 10 (Table 3). This means 467 compounds had an efflux ratio above 1 and below 10 and thus could not be classified according to the criteria. These molecules were not used to generate predictive models of efflux class. It is often the case that molecules with high efflux ratios are substrates of efflux transporters. On the other hand, molecules with very low efflux ratios are substrates of uptake transporters but there were no molecules with very low efflux ratios (< 0.1) in the caco-2 dataset. If the main mechanism by which a compound crosses a biological membrane is via passive diffusion, it is expected that the permeability of the compound in both directions will be roughly equal. With this in mind, it is expected that passive diffusion is the main permeation mechanism for low efflux compounds in this dataset. We also hypothesize that high efflux molecules are substrates of efflux transporters. Because physicochemical descriptors have a tendency to describe the likelihood of a molecule undergoing passive diffusion, it is expected that reliable descriptors of carrier-mediated transport will improve the performance of predictions of caco-2 efflux classification carrier-mediated processes are explicitly taken into account.

Table 4: Number of compounds in each of the efflux classes when compounds are clustered

High efflux	Low efflux	Total
109	252	361

Clustering the compounds and taking the centroid molecule results in a total of 361 cluster centroids, 109 of which are high efflux and 252 are low efflux centroids (Table 4).

4.2 A-B Permeability

4.2.1 Results and Discussion

From the original pool of descriptors, backward feature elimination was used to select the best combination of descriptors for Logistic regression. 31 descriptors were selected including lipophilicity (logD at pH =7.4), prototation (pKa) and MACCS keys descriptors.

Table 5: Performance of Logistic regression classifier in A-B permeability classification

Descriptors	Average Accuracy	Accuracy (sd)	P-value	Average Kappa	Kappa (sd)	P-value
Standard descriptors	0.91	0.01	-	0.79	0.01	-

The backward feature selection process did not select any of the active transport descriptors (NMTS or PIDGIN). Given that 91% of compounds were classified correctly, the combination of backward feature selection with logistic regression can be considered an effective method of selecting from a pool of descriptors, the best combination for predicting whether compounds are permeable or non-permeable in the A-B direction.

The Cohen's Kappa value of 0.79 suggests relatively good agreement between actual class and predicted class. The difference in classification accuracy and Cohen's Kappa is most likely due to the slight imbalance in the A-B permeability dataset. If the high classification accuracy is indeed due to the performance of the model, the Cohen's

kappa value could possibly be increased by balancing the datasets in terms of number of permeable and non-permeable compounds.

Compared with findings from similar studies such as by Pham The et al. (Pham The et al., 2011), this classification performance is good. However, direct comparisons cannot be made due to the differences in methodologies. For example, the datasets used in other studies contain data obtained from different sources and thus likely to contain structurally diverse sets of compounds. The learning algorithms used are also different; one advantage of such models is that they may be applicable to a wider range of molecules. However, the datasets used in such studies are often small (largest one from Pham The et al. (Pham The et al., 2011) contains 567 compounds) compared to the dataset used in this study (969 compounds).

The fact that none of the active transport descriptors were included by backward elimination suggests that *NMTS* and *PIDGIN* descriptors do not provide new information to the models. As a result, it was not possible to compare active transport descriptor based models to standard descriptor based models using the logistic regression classifier. However, it is worth bearing in mind that MACCS keys are included in the set of 'standard' descriptors. Some of the substructures encoded by MACCS keys are relevant to membrane transport. Because of the 'promiscuous' nature of membrane transporters (Kell et al., 2013), it is possible that these substructures do in fact encode the likelihood of molecules undergoing active transport. Descriptors such as the polar surface area (PSA) of the molecule are also directly related to some of the substructures such as number of oxygens in carbonyl groups. The lipophilicity of the molecule is also closely related to these substructures. After all, it is the combination of such structures in a molecule that gives rise to molecular properties such as lipophilicity and polar surface area.

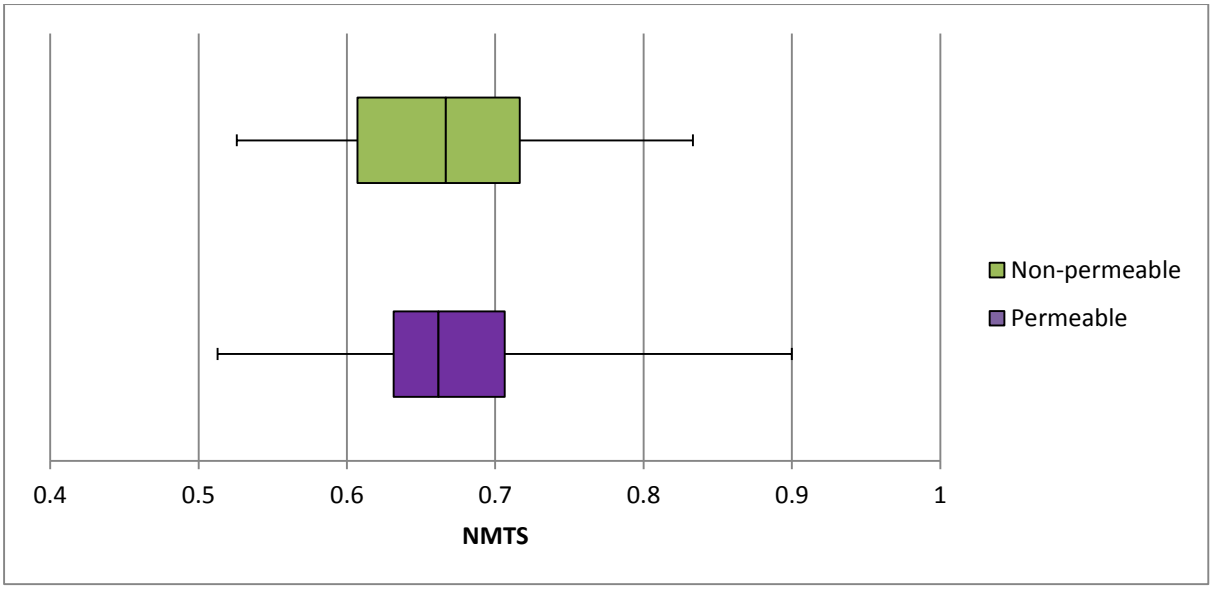

Figure 17: Distribution of NMTS values for Permeable and Impermeable compounds in the A-B direction

There was no clear distinction between the nearest metabolite Tanimoto similarity (NMTS) values of permeable and impermeable compounds in the A-B direction (Figure 17). This suggests a lack of a relationship between permeability in the A-B direction and similarity to endogenous metabolites. Interestingly, a very recent article published by the same group who suggested metabolite likeness as a measure of the likelihood of molecules undergoing carrier mediated transport shows results that agree with this study. A weak positive correlation of 0.156 (R^2) suggests no relationship between caco-2 permeability and metabolite likeness for the dataset used in that study (O'Hagan & Kell, 2015). It is likely that this is the reason why NMTS does not improve the performance of models and thus was not selected by backward feature selection.

Table 6: Important descriptors that contribute significantly to the overall accuracy of logistic regression in caco-2 Apical to Basolateral classification. The non-randomised Classification Accuracy is 91% ± 1% (0.91 ± 0.01). There were 31 descriptors to begin with. Bonferroni's critical value was 0.05/31 = 1.60E-03 (0.0016). Only descriptors with p-values below this value are shown in this table.

Randomised descriptor	**description**	**Δ mean accuracy (%)**	**P-value**
logD at pH=7.4	LogD at pH of 7.4	-3.28	9.34E-21
#XN where coord. # of X>=3	Nitrogen bonded to atom with at least 3 bonds	-2.74	2.43E-19
#ring atoms	Number of ring atoms	-1.50	9.30E-11
#S in double/charge separated bonds	Sulphur in double/charge separated bonds	-1.26	6.29E-09
#OH groups	Number of OH groups	-1.19	1.41E-08
#halogens	Number of halogens	-0.93	8.25E-07
(key(160)-1 if key(160)>1; else 0) Key160 = #CH3 groups	Number of methyl groups subtract 1	-1.12	9.05E-07
#O in C=O	Number of Oxygens in C=O groups	-0.97	1.57E-06
#N	Number of Nitrogens	-0.99	2.41E-06
#N in double bonds	Number of double bonded Nitrogens	-0.88	1.01E-05
Strongest BASIC pKa 2	pKa of the second most basic group	-0.80	1.55E-05
#N non-ring bonded to a ring	Non ring Nitrogens bonded to a ring atom	-0.68	9.91E-05
#N separated by 4 bonds	Nitrogens separated by 4 bonds	-0.69	1.42E-04
Key(164)-1 if key(164)>1; else 0 Key164 = # oxygens	Number of Oxygens subtract 1	-0.62	2.19E-04
#heteroatoms in 5 ring	Number of heteroatoms in 5 membered ring	-0.68	4.98E-04

Out of the 31 descriptors selected by backward feature selection (Appendix Table A. 2), only 16 reduced the accuracy significantly when randomised. An Independent samples t-test was carried out to compare the mean when each descriptor in the dataset is randomised to the mean when none of the descriptors are randomised. Bonferroni's method was applied to account for the multiple comparisons problem. Only those descriptors with a p-value less than the Bonferroni's critical value (calculated as 0.05/n where n is the number of comparisons considered) are shown.

The negative value for the t-statistic shows that the accuracy when each of these descriptors is randomised is less than that of the non-randomised accuracy. The magnitude of the t-value corresponds to the effect that each descriptor has on the overall accuracy when randomised.

Randomisation of logD resulted in the biggest reduction in accuracy (-3.28%). LogD represents the octanol-water distribution of ionisable compounds and is dependent on the pH of the solution(Kah & Brown, 2008). This suggests that lipophilicity is an important property for determining the permeability of compounds. Hydrogen bonding groups are widely reported to influence the membrane permeability of molecules(Refsgaard et al., 2005). The inclusion of nitrogen and oxygen containing substructures of the MACCS keys descriptor set in the list of important descriptors indicates a possible strong contribution of polarity and hydrogen bonding capacity to caco-2 permeability in the A-B direction.

That NMTS and PIDGIN descriptors are not included in the list of important descriptors does not disprove of the idea that carrier mediated transport descriptors can contribute to the overall accuracy. It is more likely that the information encoded by either nearest metabolite similarity or PIDGIN descriptors is already contained in the list of standard descriptors.

MACCS keys substructure counts were included in the list of standard descriptors. Some of these substructures (e.g. number of oxygens in carbonyl groups) are of relevance to protein binding and therefore relevant to membrane transporter binding. The fact that most of the important descriptors are MACCS keys substructures rather than physicochemical descriptors suggests that protein binding could be an important factor in determining membrane permeability. MACCS keys are often widely used to predict the bioactivity of molecules against specific targets which adds value to the idea that MACCS keys substructures may actually be encoding the likelihood of molecules undergoing active transport (Cheng et al., 2012).

One study has made use of MACCS structural keys as descriptors for development of QSAR models of permeability (Gozalbes, Jacewicz, Annand, Tsaioun, & Pineda-

Lucena, 2011). The model was developed using 14 descriptors, 9 of which are MACCS keys (Figure 18). The model shows good performance with an R^2 value of 0.77. Interestingly, some of the MACCS keys selected for model development have also been found to be important in the descriptor randomization process carried out in this study (Table 6).

Abbreviation	Descriptor definition	Coefficients
MACCS8	Heteroatoms in four-membered rings	−1.25
MACCS36	S atoms in rings	0.31
MACCS50	C in C=C bonded to >=3C	−0.27
MACCS70	N bonded to two non-C heavy atoms	0.46
MACCS100	N attached to CH_2	−0.23
MACCS119	N in double bonds	−0.28
MACCS129	CH_2s separated by 3 bonds	−0.11
MACCS132	O 2 bonds from CH_2	0.21
MACCS139	OH groups	−0.28
M Log *P*	Moriguchi octanol–water partition coefficient (Log *P*)	0.19
GATS8e	Geary autocorrelation–lag 8/weighted by atomic Sanderson electronegativities	0.11
GATS5p	Geary autocorrelation–lag 5/weighted by atomic polarizabilities	1.02
JGI1	Mean topological charge index of order 1	−4.23
JGI2	Mean topological charge index of order 2	7.43
Constant	–	−7.13

Figure 18: Descriptors and coefficients for the best QSAR model of caco-2 permeability from Gonzalbes at al. (2014)

The number of heteroatoms in four-membered rings has a coefficient of -1.25 suggesting a negative effect on permeability. In this study, it is the number of heteroatoms in five-membered ring that is important in determining whether a molecule is permeable or impermeable in the A-B direction. The number of hydroxyl (OH) groups is shown to have a negative effect on permeability. In this study, multiple models were generated making it impossible to analyse individual coefficients from individual models. The randomization process was one way to measure the importance of individual descriptors.

Substructure pattern fingerprints have been used to generate predictive classification QSAR models of human intestinal absorption and blood brain barrier permeability(Shen, Cheng, Xu, Li, & Tang, 2010). The models generated showed 94% accuracy in predicting human intestinal absorption and 69.5% accuracy for blood brain barrier permeability. It is also reported that certain substructures correlate very well with membrane permeability.

One study applied the molecular interaction field technique for predicting inhibitors of the permeability glycoprotein (P-gp) (Broccatelli et al., 2011). The study found that conformational as well as hydrophobic and hydrogen bond acceptor characteristics were important for predicting inhibitors of P-gp. This suggests a limitation in the findings of this study. The structural keys found to be important for predicting permeability may be important for both inhibitors and substrates of membrane transporters.

Certain substructures have been reported to be more prevalent in bioactive compounds and the presence of such structures can significantly increase bioactivity (Klekota & Roth, 2008). Given that a number of MACCS keys are considered important for permeability, the idea of prevalent substructures could also be relevant to membrane permeability. One would expect the prevalence of certain substructures to be more relevant to carrier mediated transport. However, it is the presence or absence of such substructures that give rise to a molecule's physicochemical properties. Such properties are therefore important for both passive diffusion carrier mediated transport processes. This highlights one of the difficulties in separating passive diffusion and carrier mediated transport process descriptors for the purpose of developing QSAR models of permeability.

Models for predicting the permeability of a group of flavonoid molecules in caco-2 cells based on both 2D and 3D descriptors show that electronic, topological, hydrogen bonding and hydrophobic properties are important determinants of permeability (Gonzales et al., 2015). This is in agreement with the results obtained in the descriptor randomization process which suggest the importance of lipophilicity (logD at pH=7.4), protonation (Strongest Basic pKa 2) and certain MACCS structural keys.

Table 7: Performance of Naïve Bayes learner in Apical to Basolateral classification

Descriptors	Average Accuracy	Accuracy (sd)	P-value (Accuracy)	Average Kappa	Kappa (sd)	P-value (Kappa)
Standard	0.83	0.01	-	0.60	0.01	-
Standard and NMTS	0.83	0.01	0.47	0.60	0.01	0.38
Standard and PIDGIN	0.83	0.00	0.46	0.60	0.01	0.47

The Naïve Bayes classifier achieved a mean accuracy of 0.83 (± 0.01) and a mean Kappa of 0.60 (± 0.01) when modelled with standard descriptors alone (Table 7). A mean accuracy of 0.83 ± 0.01 and mean Kappa of 0.60 ± 0.01 was achieved with a combination of standard descriptors and nearest metabolite similarity (NMTS). The combination of standard descriptors and PIDGIN descriptors resulted in a mean accuracy of 0.83 ± 0.01 and a mean Kappa of 0.60 ± 0.01. The p-values obtained are greater than the threshold of 0.05 which suggests that NMTS and PIDGIN descriptors did not improve the performance of the Naïve Bayes classifier in predicting A-B permeability.

The Logistic regression classifier (Table 5) generally performs better than the Naïve Bayes classifier (Table 7). This is despite the fact that more descriptors were used in generating the Naïve Bayes models. It is often mentioned in literature that the more descriptors a classifier is given the better its performance (Dearden, Cronin, & Kaiser, 2009). It could be the case that backward feature selection is a more effective method of selecting important descriptors compared with simple variance and correlation filters. However, a review of the literature shows no clear evidence suggesting certain feature selection methods are better than others (Dash & Liu, 1997; Maldonado & Weber, 2009; Yu & Liu, 2004).

It is widely reported that the more descriptors used, the more likely it is that models generated will be over-fitted to the molecules used to train the model (Dearden et al., 2009). Logistic regression models in this case are therefore likely to be better at predicting novel molecules compared with the Naïve Bayes classifier.

Table 8: Important descriptors that contribute significantly to Apical to Basolateral permeability classification Accuracy with Naïve Bayes classifier. The non-randomised accuracy is 83% ± 0.45 (0.83 ± 0.0045). The Bonferroni critical value was 0.05/144 = 3.55E-04.

Randomised descriptor	Δ mean accuracy (%)	P-value
Polar surface area	-0.74	7.60E-05

Out of 142 descriptors, randomisation of polar surface area alone causes a statistically significant reduction in the classification accuracy of the Naïve Bayes classifier (Table 8). While the reduction in accuracy is not large (-0.74), the fact that a single variable in such a high dimensional dataset can reduce the accuracy at all is interesting. The polar surface area of a molecule represents the surface area of oxygen and nitrogen atoms within the molecule. High polar surface area has been reported to be unfavourable to the caco-2 permeability of molecules(T. J. Hou et al., 2004). Although not selected as important with the Logistic regression classifier, some aspects of polar surface area may be represented by MACCS structural keys descriptors that are selected. Examples of such structural keys include the number of oxygen and nitrogen atoms. The fact that neither metabolite similarity nor PIDGIN predictions cause reduction in accuracy when randomised suggests that they do not add information to the standard descriptors.

Table 9: Performance of Random Forest classifier in Apical to Basolateral (A-B) classification.

Descriptors	Average Accuracy	Accuracy (sd)	P-value (Accuracy)	Average Kappa	Kappa (sd)	P-value (Kappa)
Standard	0.88	0.01	-	0.72	0.02	-
Standard and NMTS	0.88	0.01	0.36	0.72	0.02	0.40
Standard and PIDGIN	0.88	0.01	0.55	0.72	0.03	0.58

Similarly to the Naïve Bayes classifier, the results of the Random Forest classifier suggest no improvement when metabolite similarity and PIDGIN predictions are added (Table 9). The same set of descriptors were used for both Random Forest and

Naïve Bayes classifiers. The Random Forest classifier however shows improved classification accuracy and Kappa compared with Naïve Bayes classifier (Table 7). This is probably due to the nature of both algorithms. One of the main limitations of the Naïve Bayes classifier is that it uses all descriptors and makes the 'naïve' assumption that all descriptors are independent of one another(Wang et al., 2012). The WEKA implementation of the Random Forest classifier selects randomly 8 descriptors to generate a tree(Hall et al., 2009). This most likely reduces the likelihood of overfitting the model and thus increases the predictive performance. One study, concerned with predictions of ADME properties, found that the Random Forest classifier is at least as good and in most cases superior to the Naïve Bayes classifier (B. Chen, Sheridan, Hornak, & Voigt, 2012). These findings are in agreement with the results obtained in this study.

None of the descriptors, when randomized, caused significant reduction in the overall A-B permeability classification accuracy. This is most likely because of the bootstrapping method applied in the Random Forest classifier. Important descriptors for model performance may vary for each bootstrap, hence, randomisation of each descriptor in turn may not cause a significant reduction in accuracy. Random Forest classifiers are widely reported to perform well with high dimensional data because of their inherent ability to ignore irrelevant descriptors (Svetnik et al., 2003). In that case, it may be possible that some descriptors in the dataset were related to each other. As a result, the randomisation of such descriptors is unlikely to reduce the predictive accuracy of the Random Forest classifier.

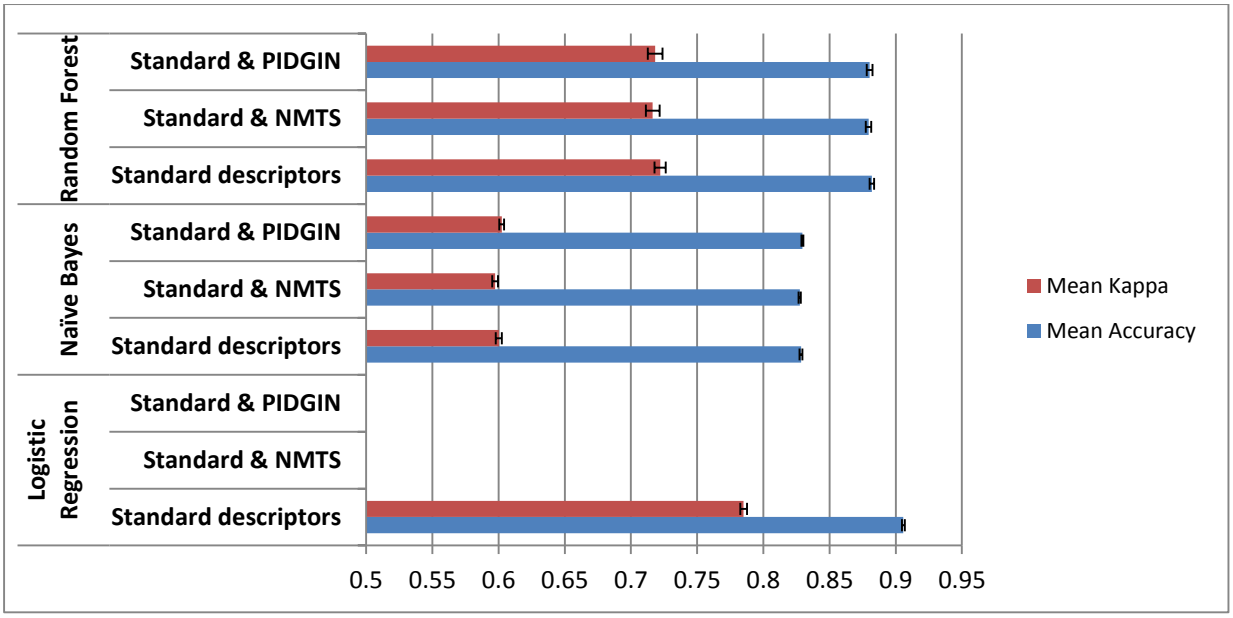

Figure 19: Performance of the different classifiers in the A-B permeability classification

Figure 19 shows that Logistic regression classification resulted in the highest classification accuracy and Cohen's Kappa values while the Naïve Bayes classifier was the worst. All the methods appear to show good classification performance of at least 80%. The results of this study are in general similar to those obtained in similar studies(Pham The et al., 2011; Refsgaard et al., 2005).

4.2.2 Conclusions

In drug discovery, the permeability of molecules is an important property for which in silico predictive models are sought. Indeed many studies have attempted to predict human intestinal permeability through data obtained from *in-vitro* models. However, there is no attempt in literature to model passive diffusion and carrier mediated transport processes separately. The main aim of this study was to investigate whether the inclusion of novel descriptors of carrier mediated transport improves the predictive performance of models. It is evident that the different methods applied result in models that perform differently in terms of overall classification accuracy and Cohen's Kappa. Statistically, novel descriptors did not improve the predictive accuracy of models of A-B permeability. The descriptor randomization process however showed the importance of certain substructures in

determining permeability. Such substructures contribute to the physicochemical properties of the molecule, suggesting contribution to passive diffusion, but may also be important in binding to membrane transporters. The combination of backward feature selection with Logistic regression classification yielded the best performance in terms of classification accuracy and Cohen's Kappa. This suggests that this method can be of use in predicting the permeability of molecules in the absorptive direction. The Random Forest classifier shows higher performance compared with the Naïve Bayes classifiers despite using the same set of descriptors for model development.

4.3 B-A Permeability

4.3.1 Results and Discussion

The permeability of a drug molecule in the B-A direction can have an impact on its efficacy. It is therefore desirable to predict whether or not molecules are permeable in the B-A direction. This section presents the performance of QSAR models of B-A permeability. Furthermore, models generated from 'standard' physicochemical and structural property descriptors are compared with models generated from a combination of physicochemical property and carrier-mediated transport descriptors.

Table 10: Performance of Logistic regression classifier in Basolateral to Apical (B-A) permeability

Descriptors	Average Accuracy	Accuracy (sd)	P-value	Average Kappa	Kappa (sd)	P-value
Standard	0.93	0.00	-	0.26	0.04	-

The backward feature selection was used for selecting descriptors for the Logistic regression classification of B-A permeability. The Logistic regression classifier shows good performance in terms of overall classification accuracy (93%) but poor performance (0.26) in terms of the Cohen's Kappa value (Table 10). This is most likely due to the imbalance in the dataset used to develop the models, which means the

probability of random agreement between predicted and actual class is high. It is reported that one of the problems associated with learning from imbalanced datasets is that the models often perform poorly on the minority class(Jeni, Cohn, De La Torre, & others, 2013).

The backward feature selection method did not select any of the novel descriptors of carrier mediated transport (NMTS and PIDGIN predictions) as important. As a result we could not compare the performance of models generated from inclusion of active transport descriptors with models generated from standard physicochemical property descriptors.

Table 11: Important descriptors that contribute significantly to B to A classification with Logistic Regression classifier. The non-randomised accuracy was 94% ± 0.37 (0.94 ± 0.0037).

Randomised descriptor	**Description**	**Δ mean accuracy (%)**	**P-value**
# O in C=O	Number of Oxygens in C=O	-0.75	4.34E-08
# 5-membered rings	Number of 5 membered rings	-0.52	1.19E-05
# ring atoms	Number of ring atoms	-0.56	8.49E-05
# O in rings	Number of Oxygen atoms in rings	-0.41	2.38E-04
# CH2s separated by 3 bonds	Number of CH2 groups separated by 3 bonds	-0.48	3.47E-04
# N	Number of Nitrogen atoms	-0.46	4.77E-04

From the descriptors selected by backward feature selection (Appendix Table A. 3), six MACCS structural keys appear to be significant contributors in prediction of B-A apparent permeability class according to descriptor randomization (Table 11). The inclusion of oxygen and nitrogen containing substructures is in agreement with the idea that B-permeability is highly driven by efflux transporters (Dolghih & Jacobson, 2013b). This is most likely due to formation of hydrogen bonds and hydrophobic interactions with efflux transporters. Indeed many studies concerned with predicting substrates of efflux transporters make use of pharmacophore models, most of which find hydrogen bond donors and acceptors as well as hydrophobic properties as important for interactions(Chang, Ekins, Bahadduri, & Swaan, 2006; Garrigues et al., 2002).

In comparison with the list of descriptors found to be important in the A-B direction, this list contains much fewer descriptors. This may suggest that not as many descriptors are relevant to the permeability of molecules in the B-A direction, a likely result of the 'promiscuous' nature of efflux transporters (Kell et al., 2013; Wong, Ma, Rothnie, Biggin, & Kerr, 2014). The biggest reduction in accuracy is caused by the randomization of number of carbonyl groups (#O in C=O). This adds more value to the idea that hydrogen bonding is an important contributor to the permeability of compounds in the B-A. In future studies, it would be ideal to have the coefficients to determine whether hydrogen bonding has a negative or positive effect on permeability. It would be also be beneficial to investigate the prevalence of such structures in permeable and non-permeable molecules.

Table 12: Performance of Naïve Bayes classifier in B-A permeability classification

Descriptors	**Average Accuracy**	**Accuracy (sd)**	**P-value**	**Average Kappa**	**Kappa (sd)**	**P-value**
Standard	0.82	0.01	-	0.15	0.02	-
Standard and NMTS	0.82	0.01	0.06	0.17	0.02	0.04
Standard and PIDGIN	0.82	0.01	0.44	0.16	0.02	0.43

The results of the Naïve Bayes classifier also suggest that the novel descriptors of carrier mediated transport do not improve the model performance (Table 12). There is a statistically significant improvement in Kappa when metabolite similarity is combined with the set of standard descriptors ($p = 0.04$). However, the difference in the Kappa values cannot be considered large in practical terms. In terms of classification accuracy, the result of including metabolite similarity is approaching the level of significance ($p = 0.06$).

None of the descriptors reduce the predictive accuracy of the Naïve Bayes classifier significantly when randomised. There are two possible reasons for this. Firstly, with 144 descriptors, it is unlikely that randomisation of a single descriptor will have much of an impact on the model performance. Secondly, the Naïve Bayes classifier makes a basic assumption that all descriptors are independent of one another.

Table 13: Performance of Random Forest classifier in B-A permeability classification

Descriptors	Average Accuracy	Accuracy (sd)	P-value	Average Kappa	Kappa (sd)	P-value
Standard	0.94	0.00	-	0.19	0.06	-
Standard and NMTS	0.94	0.00	0.64	0.19	0.05	0.73
Standard and PIDGIN	0.94	0.00	0.90	0.19	0.05	0.75

There were no statistically significant results in B-A permeability classification with the Random Forest classifier which suggests novel descriptors of carrier mediated transport failed to improve the performance of the model (Table 13). However, for Naïve Bayes and Random Forest classifiers, over 100 standard descriptors are used in model generation. It is highly unlikely that a single descriptor, nearest metabolite similarity, will make an impact when included in such a pool of descriptors.

When randomised, no individual descriptor reduces the classification accuracy of the Random Forest classifier significantly. The reasons for this are most likely similar to those mentioned above for the Naïve Bayes classifier.

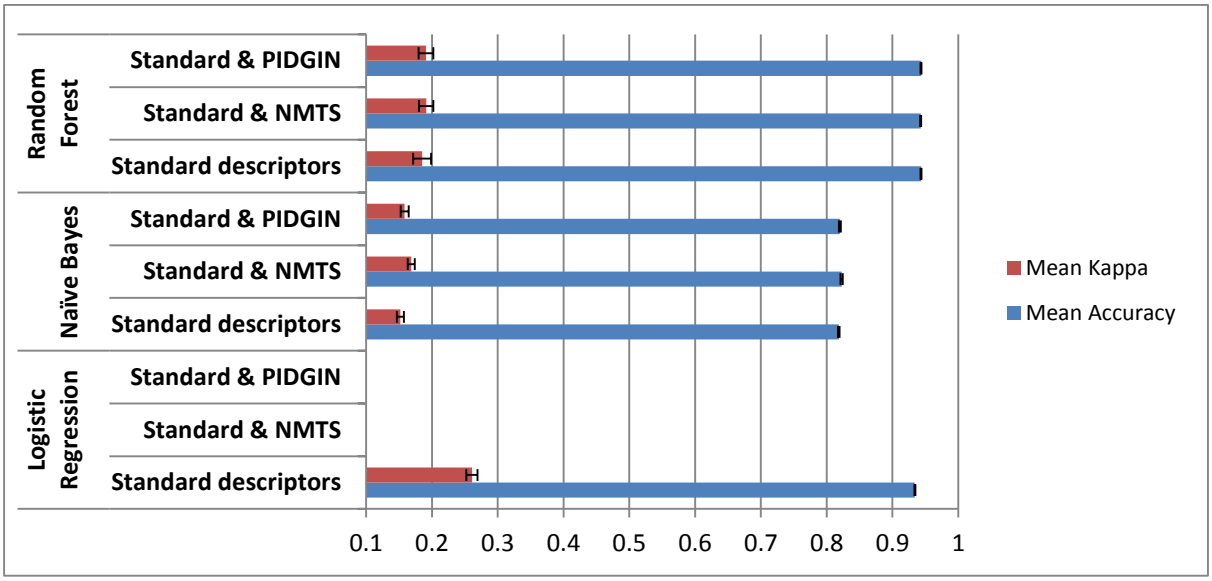

Figure 20: B-A permeability classification of all models

The Random Forest and Logistic regression classifier appear to be roughly matched in terms of classification accuracy (Figure 20). The Naïve Bayes classifier shows the

lowest classification accuracy and Cohen's Kappa values. This is an interesting finding considering the Random Forest and Naïve Bayes classifiers were generated using the same set of descriptors. This is most likely an indication that the Naïve Bayes classifier is more prone to overfitting than the Random Forest classifier. As mentioned previously, the Random Forest generates a tree from 8 randomly selected descriptors which reduces the chance of overfitting the model.

Due to the unbalanced B-A dataset used, low Cohen's Kappa values were obtained. To circumvent this, the datasets were balanced and the Naïve Bayes and Random Forest classification models were developed (Table 14 and Table 15).

The above mentioned classifiers were developed using a highly imbalanced dataset. Over 90% of the molecules belonged to the permeable set of compounds. To address this issue, the dataset was balanced and the Naïve Bayes and Random Forest classification models were developed. Because of the long running time of backward feature selection, the Logistic regression classifier was not used in this case. Balancing of datasets was carried out by splitting the permeable compounds into 14 sets, each containing 55 permeable molecules to roughly match the non-permeable class containing 50 molecules.

Table 14: Performance of the Naïve Bayes algorithm when datasets are balanced

Descriptors	**Average Accuracy**	**Accuracy (sd)**	**P-value**	**Average Kappa**	**Kappa (sd)**	**P-value**
Standard	0.83	0.09	-	0.66	0.17	-
Standard and NMTS	0.84	0.08	0.78	0.63	0.15	0.78
Standard and PIDGIN	0.82	0.09	0.48	0.67	0.17	0.48

The performance of the Naïve Bayes classifier, in terms of overall classification accuracy, does not change significantly when datasets are balanced (Table 14). For example, with standard descriptors, the average classification accuracy of an unbalanced dataset is 82% (Table 12) whereas the classification accuracy on a balanced dataset with the same descriptors is 83% (Table 14). One would expect a

reduction in the performance of the model on a balanced dataset, however, the same number of descriptors were used for the imbalanced and balanced datasets.

Table 15*:* Performance of the Random Forest classifier when datasets are balanced and cluster centroids used in model generation

Descriptors	Average Accuracy	SD Accuracy	P-value	Average Kappa	SD Kappa	P-value
Standard	0.82	0.10	-	0.64	0.19	-
Standard and NMTS	0.82	0.10	0.86	0.65	0.19	0.86
Standard and PIDGIN	0.83	0.09	0.75	0.65	0.18	0.75

The performance of Random Forest classifiers, contrary to the Naïve Bayes classifier, appears to be influenced by the imbalance in the dataset (Table 15). For example, the highest B-A permeability classification accuracy with an unbalanced dataset for the Random Forest classifier is 94% (Table 13). When the model is trained on datasets with balanced classes, the highest classification accuracy achieved is 83% (Table 15), when standard descriptors are combined with PIDGIN predictions. This finding is in agreement with studies that have made use of the Random Forest classifier and it is evidence of a possible disadvantage of the Random Forest classifier compared with the Naïve Bayes classifier. In this case, the result does not appear to be influenced by the class imbalance in the dataset.

4.3.2 Conclusions

The inclusion of novel descriptors of carrier mediated transport (*NMTS and PIDGIN predictions*) failed to improve the performance of models in B-A permeability classification. This is evidenced by the lack of p-values below the significance threshold ($p<0.05$). However, the descriptor randomisation technique employed in this study showed that some structural descriptors, e.g. number of carbonyl groups and the presence of rings, are important features that may determine the permeability of molecules in the B-A direction. While such substructures have an influence on the lipophilicity, and consequently likelihood of permeability via passive diffusion, one can argue that such substructures are also relevant to protein binding.

As a result, one can argue that the capability of molecules undergoing carrier-mediated transport is encoded in those substructures.

To date, not many QSAR models of B-A permeability have been published. The most likely reason is that the interest of drug discovery projects lies in the prediction of intestinal absorption of molecules, and as such, many studies have focused on predicting permeability in the absorptive direction (Deli et al., 2005; Dolghih & Jacobson, 2013a; Guangli & Yiyu, 2006). While the importance of predicting A-B permeability is acknowledged, the results of this study suggest that predictions of B-A permeability are possible, and as such, should be pursued. Such models may be useful in predicting, and thus help to avoid, molecules that are highly permeable in the B-A direction. Because not many studies have pursued B-A permeability predictions, it was not possible to compare results obtained in this study with other studies.

4.4 Efflux ratio classification

4.4.1 Results and Discussion

The ratio of B-A and A-B permeability (efflux ratio) is widely used to determine the extent to which permeability of molecules is affected by efflux transporters. It is therefore important to be able to predict the likelihood of molecules undergoing efflux. The aim of this section is to present the results of the predictions of classification methods.

Table 16: Performance of Logistic regression learner in Efflux ratio classification

Descriptors	Average Accuracy	Accuracy (sd)	P-value	Average Kappa	Kappa (sd)	P-value
Standard	0.88	0.01	-	0.71	0.01	-

The backward feature selection method was used to select descriptors for efflux ratio class predictions for the Logistic regression classifier (Table 16). None of the novel descriptors of carrier mediated transport were selected which suggests that

they do not add extra information to that already contained in the physicochemical and structural property descriptors. On average, 88% of the molecules were correctly classified. The Kappa value of 0.71 suggests good agreement between actual and predicted class taking random agreement into consideration. This suggests that the combination of backward feature selection and Logistic regression classification can be useful in predicting the efflux ratio of molecules. Such predictions could be useful for identifying and avoiding potential substrates of efflux transporters.

In order to identify descriptors that contribute the most to model performance, each descriptor was randomized and the resulting accuracy compared to that of the non-randomised model using the independent samples t-test.

Table 17: Important descriptors that contribute significantly to efflux classification with Logistic regression. The non-randomised accuracy is 0.88 ± 0.01. The number of descriptors selected is 9. The Bonferroni critical value applied is 5.56E-03 (0.05/9)

Randomised descriptor	**Description**	**Δ mean Accuracy (%)**	**P-value**
#N	Number of Nitrogen atoms	-3.30	2.86E-21
#non-C with coordination number >=3	Number of non-C atoms bonded to at least 3 atoms	-2.85	1.04E-19
#O in C=O	Number of oxygen in C=O groups	-2.59	3.43E-19
#N bonded to >= 3 C	Number of N bonded to 3 or more carbons	-2.43	8.39E-18
#S in double/charge separated bonds	Sulphur in double/charge separated bonds	-2.42	1.57E-16
#CH3 groups	Methyl groups	-1.98	9.53E-14
#atoms separated by (!:):(!:)	Four atoms connected by non-aromatic bonds	-1.53	5.30E-12
Strongest ACIDIC pKa 1	pKa of the strongest acidic group	-1.29	9.25E-11
#heteroatoms in 5 ring	heteroatoms in 5 membered rings	-0.63	1.66E-04

Table 17 shows the descriptors that, when randomised, cause a significant reduction in the predictive accuracy of efflux classification with the logistic regression classifier

(refer to Appendix Table A. 4 for full list of descriptors). The inclusion of nitrogen and oxygen containing substructures suggests the importance of polarity and possibly hydrogen bonding on efflux ratio. Hydrogen bond acceptor groups of substrates were found to be particularly important for the interaction with P-gp (efflux transporter) (Desai, Raub, & Blanco, 2012). This is in agreement with this study which shows that the number of carbonyl groups and the number of nitrogen atoms are important determinants of efflux ratio.

The addition of methyl groups is reported to reduce the efflux ratio of prednisolone and other glucocorticoids (Yates et al., 2003). In this study, the abundance of methyl groups is also shown to have an impact on efflux ratio. For some of the descriptors, it is not clear why they are considered important in determining the efflux ratio e.g. the number of Sulphur atoms in double/charge separated bonds. The acidity of a molecule is often reported as an important property for membrane permeability, hence, it is no surprise it is included in the list of important descriptors.

Table 18: Performance of Naïve Bayes learner in Efflux ratio classification

Descriptors	**Average Accuracy**	**Accuracy (sd)**	**P-value (Accuracy)**	**Average Kappa**	**Kappa (sd)**	**P-value (Kappa)**
Standard	0.86	0.01	-	0.67	0.02	-
Standard and NMTS	0.87	0.01	0.02	0.68	0.02	0.03
Standard and PIDGIN	0.87	0.02	0.001	0.68	0.01	0.001

Table 18 shows that with the Naïve Bayes classifier, statistically significant improvements in both accuracy and Cohen's Kappa are seen with the inclusion of novel descriptors of carrier mediated transporters. However, statistical significance does not necessarily correspond to practical significance (Kenny & Montanari, 2013). The differences observed in this study, while statistically significant, do not appear practically meaningful. However, the general performance of the Naïve Bayes classifier is good. The lack of QSAR studies aiming to predict efflux ratio means that these results cannot be compared with results in literature, and provides further evidence of the novelty of this work.

Table 19: Important descriptors that contribute significantly to efflux classification with Naïve Bayes algorithm; Bonferroni's critical value = 3.55E-04; Non-randomised Accuracy = 0.87 ± 0.008;

Randomised descriptor	Δ mean Accuracy (%)	p-value
Polar surface area	-2.33	6.12E-07

Table 19 shows that only one descriptor, Polar Surface area (PSA), reduces the predictive accuracy of the Naïve Bayes classification model of efflux ratio significantly. PSA represents the surface area of the molecule that is likely to interact with polar environments and is widely used in medicinal chemistry to optimise a molecule's ability to permeate cell membranes. One would expect high PSA to reduce the likelihood of a molecule to undergo passive diffusion. For molecules that permeate predominantly via passive diffusion, one would expect PSA to have an equal effect in both A-B and B-A permeability and thus no effect on the efflux ratio. However, for molecules that permeate via both passive diffusion and carrier mediated transport, one would expect PSA to have an effect. Judging by the number of molecules permeable in the B-A direction but not permeable in the A-B direction (therefore likely to belong to the high efflux class), it is therefore no surprise that PSA is considered an important determinant for efflux ratio of molecules in the caco-2 dataset used in this study.

Table 20: Performance of Random Forest learner in Efflux ratio classification

Descriptors	Average Accuracy	Accuracy (sd)	P-value (Accuracy)	Average Kappa	Kappa (sd)	P-value (Kappa)
Standard	0.90	0.01	-	0.75	0.02	-
Standard and NMTS	0.89	0.01	0.14	0.74	0.02	0.12
Standard and PIDGIN	0.89	0.01	0.04	0.74	0.02	0.03

Surprisingly, the inclusion of PIDGIN predictions appears to reduce the classification accuracy ($p = 0.04$) and Cohen's Kappa values ($p = 0.03$) for the Random Forest classifier (Table 20). Inclusion of metabolite similarity does not appear to have an

impact on the performance of the model in terms of both accuracy and Cohen's Kappa. However the reason we get a small p-value in this case is most likely due to very small values of standard deviation. This reduction is therefore not practically meaningful. Another reason we may get a significant reduction with Random Forests is because the classifier selects 8 descriptors for each tree. It may actually be the case that none of the PIDGIN descriptors are selected. It is difficult to analyse which descriptors are being selected for the Random Forest method because over 100 models (20 iterations multiply by 5-fold cross validations) are being generated.

None of the descriptors cause significant reduction in the performance of the Random Forest classifier when randomised. The reasons for this are most likely similar to those mentioned in earlier sections, relating to the total number of descriptors in the pool and how it is unlikely to observe significant changes in model performance when a single descriptor is randomized.

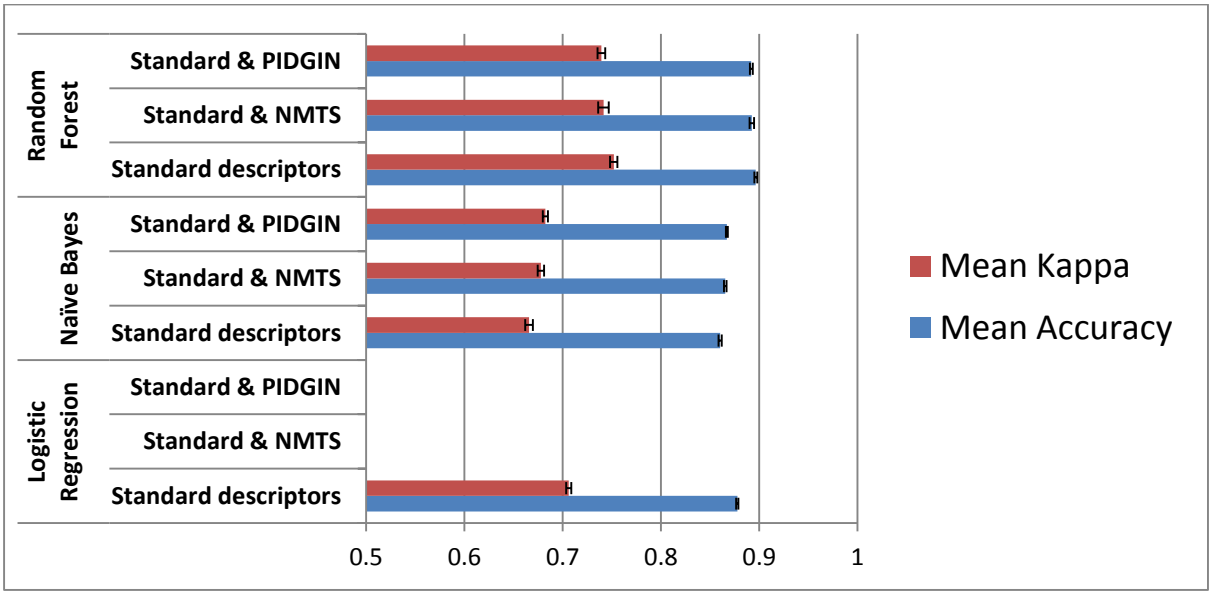

Figure 21: Performance of efflux ratio classifiers

For efflux classification with the original dataset, the Random Forest classifier shows the best performance in terms of accuracy and Kappa compared with other classification methods (Figure 21). All the methods applied result in classification accuracy of at least 85%, with the Random Forest classifier approaching a 90% correct classification rate. The lack of QSAR studies aiming to predict caco-2 efflux

ratio means that the results of this study cannot be compared to that of other studies as there is a lack of publicly available data.

Cluster centroid based classification models of efflux ratio

When molecule cluster centroids (refer to methods section 5.2) are used in model development, one expects the accuracy and kappa values of subsequent models to reduce significantly. This process eliminates the possibility of having an overrepresented chemical series in the dataset which can result in model overfitting. Such models are expected to give a more accurate picture of the likely performance of the models on external datasets.

Table 21: Performance of Logistic Regression learner in Efflux ratio classification when compounds are clustered and cluster centroid used in model development

Descriptors	**Average Accuracy**	**Accuracy (sd)**	**P-value (Accuracy)**	**Average Kappa**	**Kappa (sd)**	**P-value (Kappa)**
Standard	0.81	0.01	-	0.51	0.01	-
Standard and PIDGIN	0.85	0.01	1.43E-03	0.62	0.02	1.21E-03

After taking the cluster centroids, the backward feature selection process was carried out in the presence and in the absence of novel descriptors of carrier mediated transport. The best set of physicochemical and structural descriptors resulted in 81% of molecules correctly classified (Table 21). The models were generated from 7 descriptors, all of which were MACCS structural keys. In the presence of novel descriptors, 27 descriptors are selected by backward feature selection. The combination of standard and PIDGIN predictions resulted in 85% correctly classified molecules (Table 21). There is therefore a statistically significant increase in model performance. By reducing the possibility of having a majority chemical series within the dataset, the novel descriptors are found to be important in predicting efflux ratio. This may be due to the original dataset containing molecules belonging to the same cluster, for which PIDGIN predictions are not important determinants of efflux ratio.

One may however argue that the improvement, although statistically significant, is not very meaningful in a practical sense.

Table 22: Performance of Naïve Bayes in cluster centroid based efflux classification

Descriptors	Average Accuracy	Accuracy (sd)	P-value (Accuracy)	Average Kappa	Kappa (sd)	P-value (Kappa)
Standard	0.76	0.01	-	0.44	0.03	-
Standard and NMTS	0.76	0.01	0.81	0.44	0.02	0.49
Standard and PIDGIN	0.77	0.01	0.19	0.47	0.02	0.01

Table 22 shows the performance of the Naïve Bayes classifier in cluster centroid based efflux classification. A p-value of 0.01 indicates a significant improvement in Kappa when standard descriptors are combined with PIDGIN predictions. However, the improvement in Kappa does not appear to be large enough to have practical significance. As expected, the general performance of the classifier is reduced in comparison to when the original dataset is used. This is most likely a clearer picture of how the models are likely to perform on an external, diverse set of molecules.

Table 23: Performance of Random Forest in cluster centroid based efflux classification

Descriptors	Average Accuracy	Accuracy (sd)	P-value (Accuracy)	Average Kappa	Kappa (sd)	P-value (Kappa)
Standard	0.84	0.01	-	0.59	0.03	-
Standard and NMTS	0.84	0.01	0.49	0.59	0.03	0.52
Standard and PIDGIN	0.84	0.01	0.25	0.61	0.04	0.19

Table 23 shows the performance of Random Forest classifier in centroid based efflux classification. The lack of p-values below the threshold 0.05 suggests that NMTS and PIDGIN predictions do not improve the overall predictive accuracy or agreement between actual and predicted efflux class. However, in comparison with the non-clustered dataset, the reduction in predictive accuracy is most likely a signal that a

more representative set of molecules is used in model development. As such, one would expect this method to be more applicable to an external dataset.

4.4.2 Conclusions

Efflux ratio is used in drug discovery and other biological fields to measure the effect of efflux transport on absorption. It is therefore important make predictions of efflux ratio in order to identify and avoid substrates of efflux transporters. The lack of studies aiming to develop QSAR models of efflux ratio means that direct comparisons of the findings of this study could not be made with findings from other studies. One reason is the lack of publicly available data from which models can be developed. Most in silico studies aim to predict the A-B caco-2 permeability of molecules and as such, data on B-A permeability from which efflux ratio can be calculated is scarce. The majority of studies into efflux transport aim to develop predictive models of specific transporters. The permeability glycoprotein (P-gp) is one of the most extensively investigated efflux transporters due to its ubiquitous impact on efflux transport.

However, the classification QSAR models obtained in this study demonstrate that predictions of efflux ratio are at least as accurate as predictions of A-B and B-A permeability. This presents a cheap and fast in-silico method that can be useful in predicting molecules that are likely to undergo efflux transport prior to synthesis.

The novel descriptors did not appear to have a major impact on the performance of models in predicting efflux class. This raises questions over the usage of metabolite similarity as a metric for assessing the likelihood of molecules binding to membrane transporters. Interestingly, a recent study by O'Hagan and Kell (O'Hagan & Kell, 2015) who proposed metabolite similarity to quantify likelihood of molecules undergoing active transport, found no relationship between metabolite similarity and caco-2 permeability. Their explanation for this finding was that for some molecules, a few transporters may be relevant while for other molecules many transporters can contribute to their permeability.

The use of predictions from the PIDGIN tool could be improved by higher coverage of membrane transporters in the tool. However, the performance of models

developed from cluster centroids gives a better picture of how models are likely to perform on chemically diverse external datasets.

5. SUMMARY AND FUTURE WORK

The aim of this section is to summarise the overall conclusions of this study, highlight interesting questions that arose and give an overview of avenues that can be pursued in future studies.

5.1 Summary of conclusions

To address the relative importance of carrier mediated transport and passive diffusion through QSAR models, general descriptors of carrier mediated transport were required. Current state of the art QSAR models make use of physicochemical and structural descriptors that measure the likelihood of molecules undergoing passive diffusion. Many studies agree that current QSAR methods cannot fully account for the permeability of substrates of membrane transporters.

The aim of this study was to develop predictive classification models of caco-2 permeability that include novel descriptors of carrier-mediated transport and to evaluate the predictive performance of such models by means of statistical comparisons to models developed from standard physicochemical and structural property descriptors. In order to achieve these aims, three specific objectives were devised to guide this study.

The first objective was to assess the applicability of metabolite-likeness as a potential descriptor of the likelihood of molecules undergoing carrier-mediated transport across the cell membrane. After the proposal that structural similarity to endogenous metabolites can be used to measure the likelihood of molecules to undergo carrier mediated transport, the concept of metabolite likeness was investigated. The findings of this study suggest that approved drugs are generally more similar to endogenous metabolites than molecules found in the general chemical space considered in drug discovery projects. The degree of similarity to metabolites was found to differ with different fingerprint encodings. Both approved drugs and commercial compounds showed highest similarity to metabolites when

structures were encoded by fragment dictionary based fingerprints compared with other fingerprint encodings. In all fingerprint encodings, approved drugs were consistently more similar to endogenous metabolites than commercially available compounds. This led to the conclusion that approved drugs are more likely to bind to membrane transporters which transport the drug's structurally nearest metabolite. The nearest metabolite similarity was consequently used as a descriptor of carrier mediated transport in this study and thus the first objective was achieved.

The second objective was to assess the applicability of a target prediction tool, PIDGIN, to predict potential substrates of caco-2 expressed membrane transporters within the dataset of caco-2 tested molecules provided by Evotec. In order to use PIDGIN for the purposes of this study, an investigation was conducted to assess the abundance of human transporter models within it. Because permeability data from caco-2 cells was used, one needs to know which transporters are sufficiently expressed in caco-2 cells and have models available in the tool. A caco-2 gene expression profile was obtained and a minimum expression threshold was sought. The finding of the study was that six membrane transporters that are sufficiently expressed in caco-2 cells had models available in the target prediction tool. This is a relatively small number of membrane transporters considering many studies suggest hundreds of membrane transporters are expressed in caco-2 cells. One of the difficulties faced in this investigation is the selection of a minimum expression threshold. Many studies concerning gene expression aim to assess differential expression between cells of interest. Because of this, one may consider the approach used in this study to be novel. The PIDGIN tool was consequently used as a source of novel descriptors. The minimum protein expression threshold was double the median expression value. The PIDGIN tool was consequently used to predict substrates of caco-2 expressed membrane transporters within the dataset of caco-2 tested compounds. The binary output (0 meaning non-substrate and 1 meaning substrate of the relevant transporter) were consequently used as novel descriptors of carrier-mediated transport thus meeting the second objective of this study.

The third and final objective of this study was to develop classification models of caco-2 permeability that incorporate the novel descriptors and to evaluate whether

the inclusion of novel descriptors offers statistically significant improvements in the predictive performance of such models. To achieve this objective, classification QSAR models were developed for predicting the caco-2 cell permeability and efflux ratio class of molecules. Three classification methods were used in this study: Logistic regression, Naïve Bayes and Random Forest classifiers. Models were developed to predict the caco-2 apparent permeability (P_{app}) in the apical to basolateral (A-B) and basolateral to apical (B-A) directions as well as efflux ratio class of molecules.

The backward feature selection method was employed to select descriptors for the Logistic regression classifier. For predicting the A-B permeability, none of the novel descriptors of carrier mediated transport were selected as important. The Logistic regression classifier achieved a mean classification accuracy of 91% when generated with the selected set of physicochemical and structural property descriptors. For the Naïve Bayes and Random Forest classifiers, feature selection was carried out using low variance and correlation filters. There was no significant difference in performance between models generated from physicochemical and structural property descriptors alone and in combination with novel descriptors of carrier mediated transport. This suggests that the novel descriptors of carrier mediated transport do not improve the performance of models. To investigate the descriptors that contribute the most to predictive performance, descriptor randomization was carried out. The importance of certain substructures from the MACCS keys list suggests that the presence of some substructures in molecules could have an influence on their permeability. These substructures include hydroxyl and carbonyl groups which are relevant to protein binding. It is therefore likely that by including MACCS keys as descriptors, the likelihood of molecules undergoing carrier mediated transport is encoded. The third and final objective was therefore met since novel descriptors of carrier-mediated transport were used to develop models.

The aim of this study, to incorporate novel descriptors of carrier-mediated transport in predictive modelling of caco-2 permeability, was achieved. Structural similarity to metabolites (NMTS) and PIDGIN target predictions were used as novel descriptors of carrier-mediated transport. However, the predictive performance of models developed with novel descriptors did not show statistically significant improvements

compared with models developed without the novel descriptors. These results suggest that better descriptors of carrier-mediated descriptors should be pursued if indeed carrier-mediated transport is as important as passive diffusion to the permeability of molecules across biological membranes.

The permeability of molecules in the B-A direction may have an influence on its permeability. A review of the literature suggested that not many QSAR models of B-A permeability have been developed. For the dataset used, the majority of compounds were highly permeable in the B-A direction. It was therefore assumed that carrier-mediated transport was more prevalent in the B-A direction for the set of molecules used in this study. The dataset was imbalanced for the B-A permeability and this had an impact on the performance of the models as shown by low values of Cohen's Kappa. Only 50 molecules were impermeable in the B-A direction and more than 700 were permeable. The dataset was balanced by clustering permeable compounds and taking the centroid of each cluster for model generation. The novel descriptors of carrier mediated transport (NMTS and PIDGIN target predictions) did not improve the performance of models. The most likely reason is that the information encoded by these novel descriptors is already contained within the physicochemical and structural descriptors used. Indeed the descriptor randomization process showed that certain substructures had an influence on the B-A permeability of molecules. Such substructures, as mentioned before, are relevant to protein binding and could in fact be capturing the likelihood of molecules undergoing carrier mediated transport.

One reason why molecules may fail to show optimal permeability is because they are substrates of efflux transporters. The efflux ratio is used to measure the extent to which molecules undergo efflux transport. Because carrier mediated transport is of direct relevance to efflux ratio, one would expect novel descriptors of carrier mediated transport to be particularly important in predicting efflux ratio class. The Random Forest classifier shows the highest classification accuracy (90%) when generated from physicochemical and structural property descriptors. Addition of novel descriptors of carrier mediated transport failed to show significant improvements in the performance of classifiers on the original dataset. Clustering of

molecules in the efflux ratio dataset led to PIDGIN predictions being selected by the backward feature selection method. While models developed from the clustered dataset show statistically significant improvements with inclusion of PIDGIN predictions, the improvements were not considered practically significant.

5.2 Future work

In this study, the concept of metabolite likeness was applied to quantify the likelihood of molecules undergoing carrier mediated transport. However, the similarity to only the structurally closest endogenous metabolite was considered. In future, it may be beneficial to assess the number of endogenous metabolites to which a particular molecule is structurally very similar. For example, one may consider using the number of endogenous metabolites to which a molecule has a similarity score greater than a chosen threshold. This is likely to give a more comprehensive measure of metabolite similarity and a more robust quantification of likelihood of molecules to interact with membrane transporters.

Another avenue for future work concerns the target prediction tool used in this study. The tool was developed based on active and inactive compounds for each of the protein targets available. For membrane transporters, active molecules can be substrates or inhibitors. In future work concerned with developing target prediction tools, active molecules should be further divided into substrates or inhibitors. It would also be beneficial, in future work, to develop robust methods of selecting a minimum expression threshold above which one can be confident that a gene of interest is sufficiently expressed.

While cross validation was carried out for classification models developed in this study, in future, it is necessary to assess the performance of such models on an external dataset with compounds not used in the model development process. The performance of models on such datasets would provide a complete validation from which coherent conclusions can be made on the utility of models developed in this study. However, the lack of data on B-A permeability and thus efflux ratio could hinder the process of validating the models developed for predicting such properties.

6. REFERENCES

Anderson, J. R., Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (Eds.). (1983). *Machine learning: an artificial intelligence approach*. Los Altos, Calif: M. Kaufmann.

Arnott, J. A., & Planey, S. L. (2012). The influence of lipophilicity in drug discovery and design. *Expert Opinion on Drug Discovery*, *7*(10), 863–875. http://doi.org/10.1517/17460441.2012.714363

Ben-David, A. (2008). About the relationship between ROC curves and Cohen's kappa. *Engineering Applications of Artificial Intelligence*, *21*(6), 874–882. http://doi.org/10.1016/j.engappai.2007.09.009

Bermejo, M., Avdeef, A., Ruiz, A., Nalda, R., Ruell, J. A., Tsinman, O., … Merino, V. (2004). PAMPA—a drug absorption in vitro model. *European Journal of Pharmaceutical Sciences*, *21*(4), 429–441. http://doi.org/10.1016/j.ejps.2003.10.009

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., … Wiswedel, B. (2009). KNIME-the Konstanz information miner: version 2.0 and beyond. *AcM SIGKDD Explorations Newsletter*, *11*(1), 26–31.

Bikadi, Z., Hazai, I., Malik, D., Jemnitz, K., Veres, Z., Hari, P., … Mao, Q. (2011). Predicting P-Glycoprotein-Mediated Drug Transport Based On Support Vector Machine and Three-Dimensional Crystal Structure of P-glycoprotein. *PLoS ONE*, *6*(10), e25815. http://doi.org/10.1371/journal.pone.0025815

Borota, A., Mracec, M., Gruia, A., Rad-Curpăn, R., Ostopovici-Halip, L., & Mracec, M. (2011). A QSAR study using MTD method and Dragon descriptors for a series of selective ligands of α2C adrenoceptor. *European Journal of Medicinal Chemistry*, *46*(3), 877–884. http://doi.org/10.1016/j.ejmech.2010.12.026

Broccatelli, F., Carosati, E., Neri, A., Frosini, M., Goracci, L., Oprea, T. I., & Cruciani, G. (2011). A Novel Approach for Predicting P-Glycoprotein (ABCB1) Inhibition Using Molecular Interaction Fields. *Journal of Medicinal Chemistry*, *54*(6), 1740–1751. http://doi.org/10.1021/jm101421d

Butina, D. (1999). Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. *Journal of Chemical Information and Modeling*, *39*(4), 747–750. http://doi.org/10.1021/ci9803381

Cabrera, M. A., González, I., Fernández, C., Navarro, C., & Bermejo, M. (2006). A topological substructural approach for the prediction of P-glycoprotein substrates. *Journal of Pharmaceutical Sciences*, *95*(3), 589–606. http://doi.org/10.1002/jps.20449

Calcagno, A. M., Ludwig, J. A., Fostel, J. M., Gottesman, M. M., & Ambudkar, S. V. (2006). Comparison of Drug Transporter Levels in Normal Colon, Colon Cancer, and Caco-2 Cells: Impact on Drug Disposition and Discovery. *Molecular Pharmaceutics*, *3*(1), 87–93. http://doi.org/10.1021/mp050090k

Cereto-Massagué, A., Ojeda, M. J., Valls, C., Mulero, M., Garcia-Vallvé, S., & Pujadas, G. (2015). Molecular fingerprint similarity search in virtual screening. *Methods*, *71*, 58–63. http://doi.org/10.1016/j.ymeth.2014.08.005

Chang, C., Ekins, S., Bahadduri, P., & Swaan, P. W. (2006). Pharmacophore-based discovery of ligands for drug transporters. *Advanced Drug Delivery Reviews*, *58*(12-13), 1431–1450. http://doi.org/10.1016/j.addr.2006.09.006

Chemical Computing Group - Citing MOE. (n.d.). Retrieved 2 March 2015, from http://www.chemcomp.com/Research-Citing_MOE.htm

Chen, B., Sheridan, R. P., Hornak, V., & Voigt, J. H. (2012). Comparison of Random Forest and Pipeline Pilot Naïve Bayes in Prospective QSAR Predictions. *Journal of Chemical Information and Modeling*, *52*(3), 792–803. http://doi.org/10.1021/ci200615h

Cheng, F., Li, W., Zhou, Y., Shen, J., Wu, Z., Liu, G., … Tang, Y. (2012). admetSAR: A Comprehensive Source and Free Tool for Assessment of Chemical ADMET Properties. *Journal of Chemical Information and Modeling*, *52*(11), 3099–3105. http://doi.org/10.1021/ci300367a

Cheng, F., Zhou, Y., Li, J., Li, W., Liu, G., & Tang, Y. (2012). Prediction of chemical–protein interactions: multitarget-QSAR versus computational chemogenomic methods. *Molecular BioSystems*, *8*(9), 2373. http://doi.org/10.1039/c2mb25110h

Chen, L., Yao, J., Yang, J., & Yang, J. (2005). Predicting MDCK cell permeation coefficients of organic molecules using membrane-interaction QSAR analysis1. *Acta Pharmacologica Sinica*, *26*(11), 1322–1333. http://doi.org/10.1111/j.1745-7254.2005.00166.x

Cingolani, H. E., & Ennis, I. L. (2007). Sodium-Hydrogen Exchanger, Cardiac Overload, and Myocardial Hypertrophy. *Circulation*, *115*(9), 1090–1100. http://doi.org/10.1161/CIRCULATIONAHA.106.626929

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213.

Dash, M., & Liu, H. (1997). Feature selection for classification. *Intelligent Data Analysis*, *1*(1), 131–156.

Dearden, J. C., Cronin, M. T. D., & Kaiser, K. L. E. (2009). How not to develop a quantitative structure–activity or structure–property relationship (QSAR/QSPR). *SAR and QSAR in Environmental Research*, *20*(3-4), 241–266. http://doi.org/10.1080/10629360902949567

Dehmer, M., Varmuza, K., & Bonchev, D. (2012). *Statistical modelling of molecular descriptors in QSAR/QSPR*. Weinheim: Wiley-Blackwell. Retrieved from http://public.eblib.com/choice/publicfullrecord.aspx?p=1021397

Deli, M. A., Ábrahám, C. S., Kataoka, Y., & Niwa, M. (2005). Permeability Studies on In Vitro Blood–Brain Barrier Models: Physiology, Pathology, and Pharmacology. *Cellular and Molecular Neurobiology*, *25*(1), 59–127. http://doi.org/10.1007/s10571-004-1377-8

Desai, P. V., Raub, T. J., & Blanco, M.-J. (2012). How hydrogen bonds impact P-glycoprotein transport and permeability. *Bioorganic & Medicinal Chemistry Letters*, *22*(21), 6540–6548. http://doi.org/10.1016/j.bmcl.2012.08.059

Di, L., Artursson, P., Avdeef, A., Ecker, G. F., Faller, B., Fischer, H., … Sugano, K. (2012). Evidence-based approach to assess passive diffusion and carrier-mediated drug transport. *Drug Discovery Today*, *17*(15-16), 905–912. http://doi.org/10.1016/j.drudis.2012.03.015

Dobson, P. D., & Kell, D. B. (2008a). Carrier-mediated cellular uptake of pharmaceutical drugs: an exception or the rule? *Nature Reviews Drug Discovery*, *7*, 205–220.

Dobson, P. D., & Kell, D. B. (2008b). Carrier-mediated cellular uptake of pharmaceutical drugs: an exception or the rule? *Nature Reviews Drug Discovery*, *7*(3), 205–220. http://doi.org/10.1038/nrd2438

Dobson, P. D., Patel, Y., & Kell, D. B. (2009). 'Metabolite-likeness' as a criterion in the design and selection of pharmaceutical drug libraries. *Drug Discovery Today*, *14*(1-2), 31–40. http://doi.org/10.1016/j.drudis.2008.10.011

Dolghih, E., & Jacobson, M. P. (2013a). Predicting Efflux Ratios and Blood-Brain Barrier Penetration from Chemical Structure: Combining Passive Permeability with Active Efflux by P-Glycoprotein. *ACS Chemical Neuroscience*, *4*(2), 361–367. http://doi.org/10.1021/cn3001922

Dolghih, E., & Jacobson, M. P. (2013b). Predicting Efflux Ratios and Blood-Brain Barrier Penetration from Chemical Structure: Combining Passive Permeability with Active Efflux by P-Glycoprotein. *ACS Chemical Neuroscience*, *4*(2), 361–367. http://doi.org/10.1021/cn3001922

Durant, J. L., Leland, B. A., Henry, D. R., & Nourse, J. G. (2002). Reoptimization of MDL Keys for Use in Drug Discovery. *Journal of Chemical Information and Modeling*, *42*(6), 1273–1280. http://doi.org/10.1021/ci010132r

Eckert, H., & Bajorath, J. (2007). Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discovery Today*, *12*(5-6), 225–233. http://doi.org/10.1016/j.drudis.2007.01.011

Ekins, S., Mestres, J., & Testa, B. (2007). In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling. *British Journal of Pharmacology*, *152*(1), 9–20.

Eklund, M., Norinder, U., Boyer, S., & Carlsson, L. (2014). Choosing Feature Selection and Learning Algorithms in QSAR. *Journal of Chemical Information and Modeling*, *54*(3), 837–843. http://doi.org/10.1021/ci400573c

Faulon, J.-L., & Bender, A. (Eds.). (2010). *Handbook of chemoinformatics algorithms*. Boca Raton, FL: Chapman & Hall/CRC.

Fujita, T. (1995). *QSAR and drug design new developments and applications*. Amsterdam; New York: Elsevier. Retrieved from http://site.ebrary.com/id/10217123

Garrigues, A., Loiseau, N., Delaforge, M., Ferté, J., Garrigos, M., André, F., & Orlowski, S. (2002). Characterization of two pharmacophores on the multidrug transporter P-glycoprotein. *Molecular Pharmacology*, *62*(6), 1288–1298.

Giacomini, K. M., Huang, S.-M., Tweedie, D. J., Benet, L. Z., Brouwer, K. L. R., Chu, X., … Zhang, L. (2010). Membrane transporters in drug development. *Nature Reviews Drug Discovery*, *9*(3), 215–236. http://doi.org/10.1038/nrd3028

Goñi, F. M. (2014). The basic structure and dynamics of cell membranes: An update of the Singer–Nicolson model. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, *1838*(6), 1467–1476. http://doi.org/10.1016/j.bbamem.2014.01.006

Gonzales, G. B., Van Camp, J., Zotti, M., Kobayashi, V., Grootaert, C., Raes, K., & Smagghe, G. (2015). Two- and three-dimensional quantitative structure–permeability relationship of flavonoids in Caco-2 cells using stepwise multiple linear regression (SMLR), partial least squares regression (PLSR), and pharmacophore (GALAHAD)-based comparative molecular similarity index analysis (COMSIA). *Medicinal Chemistry Research*, *24*(4), 1696–1706. http://doi.org/10.1007/s00044-014-1241-4

Gozalbes, R., Jacewicz, M., Annand, R., Tsaioun, K., & Pineda-Lucena, A. (2011). QSAR-based permeability model for drug-like compounds. *Bioorganic & Medicinal Chemistry*, *19*(8), 2615–2624. http://doi.org/10.1016/j.bmc.2011.03.011

Guangli, M., & Yiyu, C. (2006). Predicting Caco-2 permeability using support vector machine and chemistry development kit. *J Pharm Pharm Sci*, *9*(2), 210–21.

Guha, R., & Van Drie, J. H. (2008). Structure−Activity Landscape Index: Identifying and Quantifying Activity Cliffs. *Journal of Chemical Information and Modeling*, *48*(3), 646–658. http://doi.org/10.1021/ci7004093

Halestrap, A. P., & Wilson, M. C. (2012). The monocarboxylate transporter family-Role and regulation. *IUBMB Life*, *64*(2), 109–119. http://doi.org/10.1002/iub.572

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10–18.

Hayeshi, R., Hilgendorf, C., Artursson, P., Augustijns, P., Brodin, B., Dehertogh, P., … Ungell, A.-L. (2008). Comparison of drug transporter gene expression and functionality in Caco-2 cells from 10 different laboratories. *European Journal of Pharmaceutical Sciences*, *35*(5), 383–396. http://doi.org/10.1016/j.ejps.2008.08.004

Holliday, J. D., Salim, N., Whittle, M., & Willett, P. (2003). Analysis and Display of the Size Dependence of Chemical Similarity Coefficients. *Journal of Chemical Information and Modeling*, *43*(3), 819–828. http://doi.org/10.1021/ci034001x

Hou, T. J., Zhang, W., Xia, K., Qiao, X. B., & Xu, X. J. (2004). ADME Evaluation in Drug Discovery. 5. Correlation of Caco-2 Permeation with Simple Molecular Properties. *Journal of Chemical Information and Modeling*, *44*(5), 1585–1600. http://doi.org/10.1021/ci049884m

Hou, T., Wang, J., Zhang, W., Wang, W., & Xu, X. (2006). Recent advances in computational prediction of drug absorption and permeability in drug discovery. *Current Medicinal Chemistry*, *13*(22), 2653–2667.

Iceta, R., Aramayona, J. J., Mesonero, J. E., & Alcalde, A. I. (2008). Regulation of the human serotonin transporter mediated by long-term action of serotonin in Caco-2 cells. *Acta Physiologica*, *193*(1), 57–65. http://doi.org/10.1111/j.1748-1716.2007.01793.x

Irvine, J. D., Takahashi, L., Lockhart, K., Cheong, J., Tolan, J. W., Selick, H. E., & Grove, J. R. (1999). MDCK (Madin-Darby canine kidney) cells: A tool for membrane permeability screening. *Journal of Pharmaceutical Sciences*, *88*(1), 28–33. http://doi.org/10.1021/js9803205

Jeni, L., Cohn, J. F., De La Torre, F., & others. (2013). Facing Imbalanced Data–Recommendations for the Use of Performance Metrics. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on* (pp. 245–251). IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6681438

Jin, Y., & Wang, L. (Eds.). (2009). *Fuzzy systems in bioinformatics and computational biology*. Berlin: Springer-Verlag.

Kah, M., & Brown, C. D. (2008). LogD: Lipophilicity for ionisable compounds. *Chemosphere*, *72*(10), 1401–1408. http://doi.org/10.1016/j.chemosphere.2008.04.074

Keiser, M. J., Roth, B. L., Armbruster, B. N., Ernsberger, P., Irwin, J. J., & Shoichet, B. K. (2007). Relating protein pharmacology by ligand chemistry. *Nature Biotechnology*, *25*(2), 197–206. http://doi.org/10.1038/nbt1284

Kell, D. B. (2015a). What would be the observable consequences if phospholipid bilayer diffusion of drugs into cells is negligible? *Trends in Pharmacological Sciences*, *36*(1), 15–21. http://doi.org/10.1016/j.tips.2014.10.005

Kell, D. B. (2015b). What would be the observable consequences if phospholipid bilayer diffusion of drugs into cells is negligible? *Trends in Pharmacological Sciences*, *36*(1), 15–21. http://doi.org/10.1016/j.tips.2014.10.005

Kell, D. B., Dobson, P. D., Bilsland, E., & Oliver, S. G. (2013). The promiscuous binding of pharmaceutical drugs and their transporter-mediated uptake into cells: what we (need to) know and how we can do so. *Drug Discovery Today*, *18*(5), 218–239.

Kell, D. B., Dobson, P. D., & Oliver, S. G. (2011). Pharmaceutical drug transport: the issues and the implications that it is essentially carrier-mediated only. *Drug Discovery Today*, *16*(15-16), 704–714. http://doi.org/10.1016/j.drudis.2011.05.010

Kell, D. B., & Oliver, S. G. (2014). How drugs get into cells: tested and testable predictions to help discriminate between transporter-mediated uptake and lipoidal bilayer diffusion. *Frontiers in Pharmacology*, *5*. http://doi.org/10.3389/fphar.2014.00231

Kenny, P. W., & Montanari, C. A. (2013). Inflation of correlation in the pursuit of drug-likeness. *Journal of Computer-Aided Molecular Design*, *27*(1), 1–13. http://doi.org/10.1007/s10822-012-9631-5

Kerns, E. H., & Di, L. (2008). *Drug-like properties concepts, structure design and methods: from ADME to toxicity optimization*. Amsterdam; Boston: Academic Press. Retrieved from http://site.ebrary.com/id/10251256

Khanna, V., & Ranganathan, S. (2011). Structural diversity of biologically interesting datasets: a scaffold analysis approach. *Journal of Cheminformatics*, *3*(1), 1 – 14. http://doi.org/10.1186/1758-2946-3-30

Klekota, J., & Roth, F. P. (2008). Chemical substructures that enrich for biological activity. *Bioinformatics*, *24*(21), 2518–2525. http://doi.org/10.1093/bioinformatics/btn479

Larsen, M. B., Sonders, M. S., Mortensen, O. V., Larson, G. A., Zahniser, N. R., & Amara, S. G. (2011). Dopamine Transport by the Serotonin Transporter: A Mechanistically Distinct Mode of Substrate Translocation. *Journal of Neuroscience*, *31*(17), 6605–6615. http://doi.org/10.1523/JNEUROSCI.0576-11.2011

Leach, A. R. (2001). *Molecular modelling: principles and applications*. Harlow, England; New York: Prentice Hall.

Leach, A. R., & Gillet, V. J. (2007). *An introduction to chemoinformatics*. Dordrecht; London: Springer. Retrieved from http://dx.doi.org/10.1007/978-1-4020-6291-9

Leslie, E. M., Deeley, R. G., & Cole, S. P. C. (2005). Multidrug resistance proteins: role of P-glycoprotein, MRP1, MRP2, and BCRP (ABCG2) in tissue defense. *Toxicology and Applied Pharmacology*, *204*(3), 216–237. http://doi.org/10.1016/j.taap.2004.10.012

Lipinski, C. A., Lombardo, F., Dominy, B. W., & Feeney, P. J. (2012). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, *64*, 4–17.

Li, S., Hu, P. C., & Malmstadt, N. (2011). Imaging Molecular Transport across Lipid Bilayers. *Biophysical Journal*, *101*(3), 700–708. http://doi.org/10.1016/j.bpj.2011.06.044

Maggiora, G. M. (2006). On Outliers and Activity CliffsWhy QSAR Often Disappoints. *Journal of Chemical Information and Modeling*, *46*(4), 1535–1535. http://doi.org/10.1021/ci060117s

Magrane, M., & Consortium, U. (2011). UniProt Knowledgebase: a hub of integrated protein data. *Database*, *2011*(0), bar009–bar009. http://doi.org/10.1093/database/bar009

Maldonado, S., & Weber, R. (2009). A wrapper method for feature selection using Support Vector Machines. *Information Sciences*, *179*(13), 2208–2217. http://doi.org/10.1016/j.ins.2009.02.014

Manallack, D. (2011). The pKa Distribution of Drugs: Application to Drug Discovery. In H. Trimm & W. Hunter (Eds.), *Dyes and Drugs* (pp. 80–102). Apple Academic Press. Retrieved from http://www.crcnetbase.com/doi/abs/10.1201/b13128-7

Matsson, P., Fenu, L. A., Lundquist, P., Wiśniewski, J. R., Kansy, M., & Artursson, P. (2015). Quantifying the impact of transporters on cellular drug permeability. *Trends in Pharmacological Sciences*, *36*(5), 255–262. http://doi.org/10.1016/j.tips.2015.02.009

Mervin, L. H., Afzal, A. M., Drakakis, G., Lewis, R., Engkvist, O., & Bender, A. (2015). Target prediction utilising negative bioactivity data covering large chemical space. *Journal of Cheminformatics*, *7*(1). http://doi.org/10.1186/s13321-015-0098-y

Mestres, J., & Maggiora, G. M. (2006). Putting molecular similarity into context: asymmetric indices for field-based similarity measures. *Journal of Mathematical Chemistry*, *39*(1), 107–118. http://doi.org/10.1007/s10910-005-9007-3

Nevo, Y., & Nelson, N. (2006). The NRAMP family of metal-ion transporters. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, *1763*(7), 609–620. http://doi.org/10.1016/j.bbamcr.2006.05.007

Newby, D. A. (2014). *Data mining methods for the prediction of intestinal absorption using QSAR*. University of Kent, University of Greenwich. Retrieved from https://kar.kent.ac.uk/47600/1/DNEWBY_Thesis_FINAL_Nov14.pdf

Nicolson, G. L. (2014). The Fluid—Mosaic Model of Membrane Structure: Still relevant to understanding the structure, function and dynamics of biological membranes after more than 40years. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, *1838*(6), 1451–1466. http://doi.org/10.1016/j.bbamem.2013.10.019

O'Hagan, S., & Kell, D. B. (2015). The apparent permeabilities of Caco-2 cells to marketed drugs: magnitude, and independence from both biophysical properties and endogenite similarities. *PeerJ*, *3*, e1405. http://doi.org/10.7717/peerj.1405

O'Hagan, S., Swainston, N., Handl, J., & Kell, D. B. (2014). A 'rule of 0.5' for the metabolite-likeness of approved pharmaceutical drugs. *Metabolomics*. http://doi.org/10.1007/s11306-014-0733-z

Okamura, A., Emoto, A., Koyabu, N., Ohtani, H., & Sawada, Y. (2002). Transport and uptake of nateglinide in Caco-2 cells and its inhibitory effect on human monocarboxylate transporter MCT1. *British Journal of Pharmacology*, *137*(3), 391–399. http://doi.org/10.1038/sj.bjp.0704875

Park, J. H., Carlin, K. P., Wu, G., Ilyin, V. I., Musza, L. L., Blake, P. R., & Kyle, D. J. (2014). Studies Examining the Relationship between the Chemical Structure of Protoxin II and Its Activity on Voltage Gated Sodium Channels. *Journal of Medicinal Chemistry*, *57*(15), 6623–6631. http://doi.org/10.1021/jm500687u

Patani, G. A., & LaVoie, E. J. (1996). Bioisosterism: a rational approach in drug design. *Chemical Reviews*, *96*(8), 3147–3176.

Patro, S., & Sahu, K. K. (2015). Normalization: A Preprocessing Stage. *arXiv Preprint arXiv:1503.06462*. Retrieved from http://arxiv.org/abs/1503.06462

Peironcely, J. E., Reijmers, T., Coulier, L., Bender, A., & Hankemeier, T. (2011). Understanding and Classifying Metabolite Space and Metabolite-Likeness. *PLoS ONE*, *6*(12), e28966. http://doi.org/10.1371/journal.pone.0028966

Pham The, H., González-Álvarez, I., Bermejo, M., Mangas Sanjuan, V., Centelles, I., Garrigues, T. M., & Cabrera-Pérez, M. Á. (2011). In Silico Prediction of Caco-2 Cell Permeability by a Classification QSAR Approach. *Molecular Informatics*, *30*(4), 376–385. http://doi.org/10.1002/minf.201000118

Prime-Chapman, H. M. (2004). Differential Multidrug Resistance-Associated Protein 1 through 6 Isoform Expression and Function in Human Intestinal Epithelial Caco-2 Cells. *Journal of Pharmacology and Experimental Therapeutics*, *311*(2), 476–484. http://doi.org/10.1124/jpet.104.068775

Rafter, J. A., Abell, M. L., & Braselton, J. P. (2002). Multiple comparison methods for means. *Siam Review*, *44*(2), 259–278.

Refsgaard, H. H. F., Jensen, B. F., Brockhoff, P. B., Guldbrandt, M., & Christensen, M. S. (2005). In Silico Prediction of Membrane Permeability from Calculated Molecular Parameters. *Journal of Medicinal Chemistry*, *48*(3), 805–811. http://doi.org/10.1021/jm049661n

Riniker, S., & Landrum, G. A. (2013). Open-source platform to benchmark fingerprints for ligand-based virtual screening. *Journal of Cheminformatics*, *5*(1), 1–17.

Rogers, D., & Hahn, M. (2010a). Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling*, *50*(5), 742–754. http://doi.org/10.1021/ci100050t

Rogers, D., & Hahn, M. (2010b). Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling*, *50*(5), 742–754. http://doi.org/10.1021/ci100050t

Rothfield, L. I. (Ed.). (1971). *Structure and function of biological membranes*. New York: Academic Press.

Roth, M., Obaidat, A., & Hagenbuch, B. (2012). OATPs, OATs and OCTs: the organic anion and cation transporters of the SLCO and SLC22A gene superfamilies: OATPs, OATs and OCTs. *British Journal of Pharmacology*, *165*(5), 1260–1287. http://doi.org/10.1111/j.1476-5381.2011.01724.x

Saeed, F., Salim, N., & Abdo, A. (2013). Consensus Methods for Combining Multiple Clusterings of Chemical Structures. *Journal of Chemical Information and Modeling*, *53*(5), 1026–1034. http://doi.org/10.1021/ci300442u

Saubern, S., Guha, R., & Baell, J. B. (2011). KNIME Workflow to Assess PAINS Filters in SMARTS Format. Comparison of RDKit and Indigo Cheminformatics Libraries. *Molecular Informatics*, *30*(10), 847–850. http://doi.org/10.1002/minf.201100076

Sedykh, A., Fourches, D., Duan, J., Hucke, O., Garneau, M., Zhu, H., … Tropsha, A. (2013). Human Intestinal Transporter Database: QSAR Modeling and Virtual Profiling of Drug Uptake, Efflux and Interactions. *Pharmaceutical Research*, *30*(4), 996–1007. http://doi.org/10.1007/s11095-012-0935-x

Sevin, E., Dehouck, L., Fabulas-da Costa, A., Cecchelli, R., Dehouck, M. P., Lundquist, S., & Culot, M. (2013). Accelerated Caco-2 cell permeability model for drug discovery. *Journal of Pharmacological and Toxicological Methods*, *68*(3), 334–339. http://doi.org/10.1016/j.vascn.2013.07.004

Shen, J., Cheng, F., Xu, Y., Li, W., & Tang, Y. (2010). Estimation of ADME Properties with Substructure Pattern Recognition. *Journal of Chemical Information and Modeling*, *50*(6), 1034–1041. http://doi.org/10.1021/ci100104j

Siissalo, S., Laitinen, L., Koljonen, M., Vellonen, K.-S., Kortejärvi, H., Urtti, A., … Kaukonen, A. M. (2007). Effect of cell differentiation and passage number on the expression of efflux proteins in wild type and vinblastine-induced Caco-2 cell lines. *European Journal of Pharmaceutics and Biopharmaceutics*, *67*(2), 548–554. http://doi.org/10.1016/j.ejpb.2007.03.017

Stenberg, P., Norinder, U., Luthman, K., & Artursson, P. (2001). Experimental and Computational Screening Models for the Prediction of Intestinal Drug Absorption. *Journal of Medicinal Chemistry*, *44*(12), 1927–1937. http://doi.org/10.1021/jm001101a

Stryer, L., & Gumport, R. I. (1995). *Biochemistry. - - Student's companion / Richard I. Gumport ... [et al.]. - 4. ed. - 1995. - 795 s.: ill. - 0716725606 - - Student's companion / Richard I. Gumport ... [et al.]. - 4. ed. - 1995. - 795 s. : ill. - 0716725606*. New York: W. H. Freeman and company.

Stumpfe, D., Hu, Y., Dimova, D., & Bajorath, J. (2014). Recent Progress in Understanding Activity Cliffs and Their Utility in Medicinal Chemistry: Miniperspective. *Journal of Medicinal Chemistry*, *57*(1), 18–28. http://doi.org/10.1021/jm401120g

Sugano, K., Kansy, M., Artursson, P., Avdeef, A., Bendels, S., Di, L., … Senner, F. (2010). Coexistence of passive and carrier-mediated processes in drug transport. *Nature Reviews Drug Discovery*, *9*(8), 597–614. http://doi.org/10.1038/nrd3187

Sun, D., Lennernas, H., Welage, L. S., Barnett, J. L., Landowski, C. P., Foster, D., … Amidon, G. L. (2002). Comparison of human duodenum and Caco-2 gene expression profiles for 12,000 gene sequences tags and correlation with permeability of 26 drugs. *Pharmaceutical Research*, *19*(10), 1400–1416.

Sun, H., & Pang, K. S. (2007). Permeability, Transport, and Metabolism of Solutes in Caco-2 Cell Monolayers: A Theoretical Study. *Drug Metabolism and Disposition*, *36*(1), 102–123. http://doi.org/10.1124/dmd.107.015321

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of Chemical Information and Modeling*, *43*(6), 1947–1958. http://doi.org/10.1021/ci034160g

Testa, B., Crivori, P., Reist, M., & Carrupt, P.-A. (2000). The influence of lipophilicity on the pharmacokinetic behavior of drugs: concepts and examples. *Perspectives in Drug Discovery and Design*, *19*(1), 179–211.

Tuv, E., Borisov, A., Runger, G., & Torkkola, K. (2009). Feature selection with ensembles, artificial variables, and redundancy elimination. *The Journal of Machine Learning Research*, *10*, 1341–1366.

Vaidyanathan, J. B., & Walle, T. (2001). Transport and metabolism of the tea flavonoid (–)-epicatechin by the human intestinal cell line Caco-2. *Pharmaceutical Research*, *18*(10), 1420–1425.

van der Kamp, M. W., & Mulholland, A. J. (2013). Combined Quantum Mechanics/Molecular Mechanics (QM/MM) Methods in Computational Enzymology. *Biochemistry*, *52*(16), 2708–2728. http://doi.org/10.1021/bi400215w

Volpe, D. A. (2008). Variability in Caco-2 and MDCK cell-based intestinal permeability assays. *Journal of Pharmaceutical Sciences*, *97*(2), 712–725. http://doi.org/10.1002/jps.21010

Wang, S., Li, Y., Wang, J., Chen, L., Zhang, L., Yu, H., & Hou, T. (2012). ADMET Evaluation in Drug Discovery. 12. Development of Binary Classification Models for Prediction of hERG Potassium Channel Blockage. *Molecular Pharmaceutics*, *9*(4), 996–1010. http://doi.org/10.1021/mp300023x

Waring, M. J. (2010). Lipophilicity in drug discovery. *Expert Opinion on Drug Discovery*, *5*(3), 235–248. http://doi.org/10.1517/17460441003605098

Warr, W. A. (2012). Scientific workflow systems: Pipeline Pilot and KNIME. *Journal of Computer-Aided Molecular Design*, *26*(7), 801–804. http://doi.org/10.1007/s10822-012-9577-7

Wishart, D. S. (2006). DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research*, *34*(90001), D668–D672. http://doi.org/10.1093/nar/gkj067

Wishart, D. S., Tzur, D., Knox, C., Eisner, R., Guo, A. C., Young, N., … Querengesser, L. (2007). HMDB: the Human Metabolome Database. *Nucleic Acids Research*, *35*(Database), D521–D526. http://doi.org/10.1093/nar/gkl923

Wong, K., Ma, J., Rothnie, A., Biggin, P. C., & Kerr, I. D. (2014). Towards understanding promiscuity in multidrug efflux pumps. *Trends in Biochemical Sciences*, *39*(1), 8–16. http://doi.org/10.1016/j.tibs.2013.11.002

Yates, C. R., Chang, C., Kearbey, J. D., Yasuda, K., Schuetz, E. G., Miller, D. D., … Swaan, P. W. (2003). Structural determinants of P-glycoprotein-mediated transport of glucocorticoids. *Pharmaceutical Research*, *20*(11), 1794–1803.

Yee, L. C., & Wei, Y. C. (2012). Current modeling methods used in QSAR/QSPR. *Statistical Modeling of Molecular Descriptors in QSAR/QSPR, Wiley-VCH Verlag GmbH & Co., KGaA Weinhein*. Retrieved from http://media.johnwiley.com.au/product_data/excerpt/48/35273243/3527324348-271.pdf

Yu, L., & Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *The Journal of Machine Learning Research*, *5*, 1205–1224.

7. Appendix

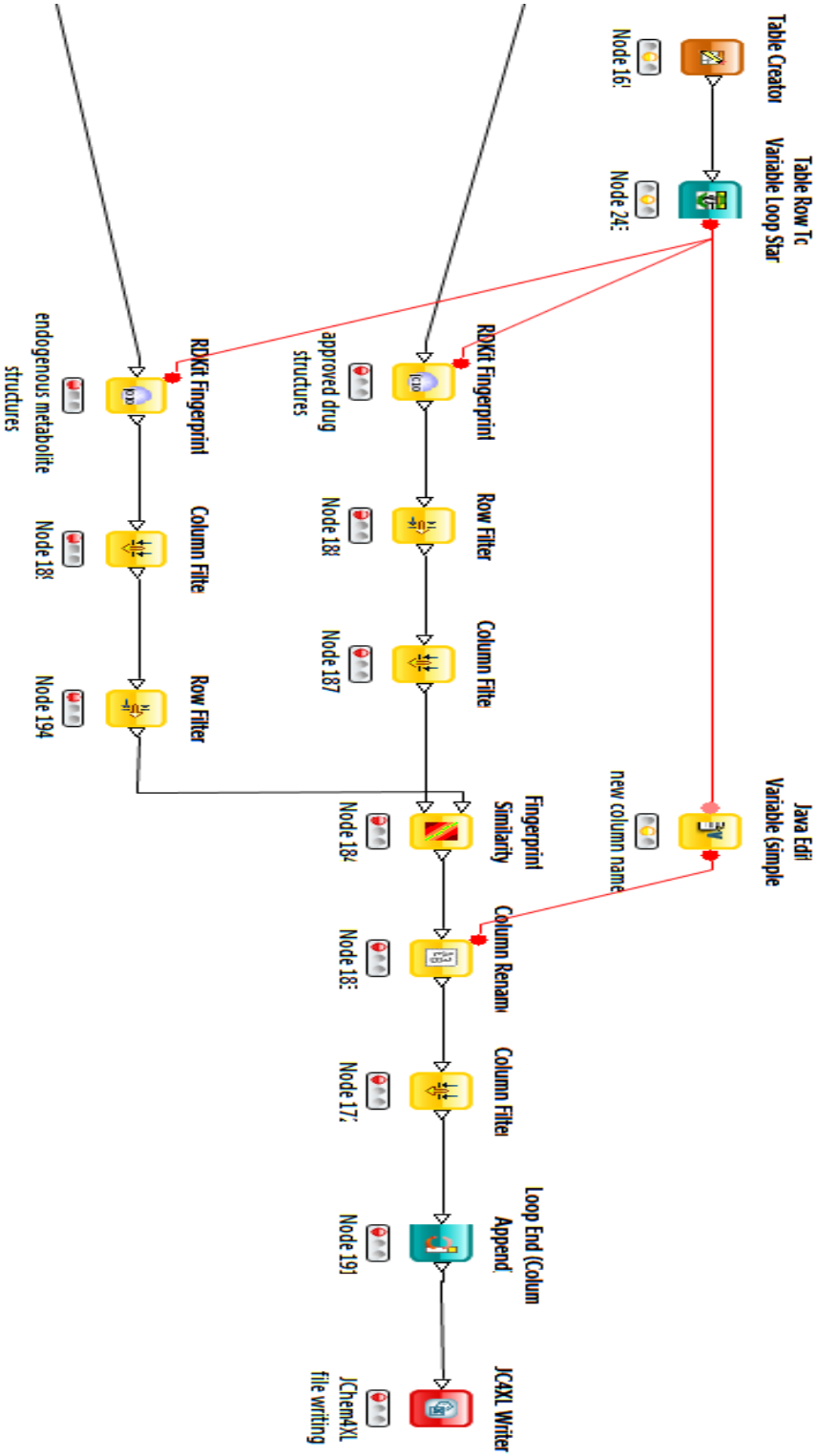

Figure A. 1: KNIME workflow used to compare metabolite-likeness of approved drugs and commercially available compounds (Evosource compounds)

Table A. 1: Proteins with models available in the target prediction tool PIDGIN and expressed in caco-2 cells

UniProt Accession	Protein name	Caco-2 RMA gene expression
P49281	solute carrier family 11 (proton-coupled divalent metal ion transporter), member 2	9.16
P04114	apolipoprotein B	9.11
Q15125	emopamil binding protein (sterol isomerase)	8.96
P53985	solute carrier family 16 (monocarboxylate transporter), member 1	7.62
P06213	insulin receptor	7.28
Q05513	protein kinase C, zeta	7.11
P13569	cystic fibrosis transmembrane conductance regulator (ATP-binding cassette sub-family C, member 7)	7.04
P19634	solute carrier family 9, subfamily A (NHE1, cation proton antiporter 1), member 1	7.01
Q8WTV0	scavenger receptor class B, member 1	6.99
P31645	solute carrier family 6 (neurotransmitter transporter), member 4	6.81
Q9UNQ0	ATP-binding cassette, sub-family G (WHITE), member 2 (Junior blood group)	6.68
P33527	ATP-binding cassette, sub-family C (CFTR/MRP), member 1	6.64
Q02156	protein kinase C, epsilon	6.64
Q9HBY8	serum/glucocorticoid regulated kinase 2	6.62
O00141	serum/glucocorticoid regulated kinase 1	6.60
P31749	v-akt murine thymoma viral oncogene homolog 1	6.55
P43003	solute carrier family 1 (glial high affinity glutamate transporter), member 3	6.49
P23975	solute carrier family 6 (neurotransmitter transporter), member 2	6.48
Q99808	solute carrier family 29 (equilibrative nucleoside transporter), member 1	6.44
P03372	estrogen receptor 1	6.16
P31639	solute carrier family 5 (sodium/glucose cotransporter), member 2	6.08
P10415	B-cell CLL/lymphoma 2	6.07
P43005	solute carrier family 1 (neuronal/epithelial high affinity glutamate transporter, system Xag), member 1	6.03
P08183	ATP-binding cassette, sub-family B (MDR/TAP), member 1	5.95
P42345	mechanistic target of rapamycin (serine/threonine kinase)	5.87

Q16602	calcitonin receptor-like	5.85
O60894	receptor (G protein-coupled) activity modifying protein 1	5.84
Q99523	sortilin 1	5.82
Q01959	solute carrier family 6 (neurotransmitter transporter), member 3	5.80
Q12908	solute carrier family 10 (sodium/bile acid cotransporter), member 2	5.66
Q16572	solute carrier family 18 (vesicular acetylcholine transporter), member 3	5.57
Q09428	ATP-binding cassette, sub-family C (CFTR/MRP), member 8	5.48
Q9Y345	solute carrier family 6 (neurotransmitter transporter), member 5	5.35
P13866	solute carrier family 5 (sodium/glucose cotransporter), member 1	5.27
P22748	carbonic anhydrase IV	5.23
O60706	ATP-binding cassette, sub-family C (CFTR/MRP), member 9	5.08
Q9NY91	solute carrier family 5 (glucose activated ion channel), member 4	5.07
P48067	solute carrier family 6 (neurotransmitter transporter, glycine), member 9	5.06
Q13255	glutamate receptor, metabotropic 1	5.02
P31644	gamma-aminobutyric acid (GABA) A receptor, alpha 5	4.96
P43004	solute carrier family 1 (glial high affinity glutamate transporter), member 2	4.95
P05771	protein kinase C, beta	4.88
P00918	carbonic anhydrase II	4.84
P31751	v-akt murine thymoma viral oncogene homolog 2	4.81
P54646	protein kinase, AMP-activated, alpha 2 catalytic subunit	4.74
P08913	adrenoceptor alpha 2A	4.32
Q9Y210	transient receptor potential cation channel, subfamily C, member 6	4.30
Q13131	protein kinase, AMP-activated, alpha 1 catalytic subunit	4.14
P30531	solute carrier family 6 (neurotransmitter transporter), member 1	3.71

RMA: Robust Multi-array Average gene expression values

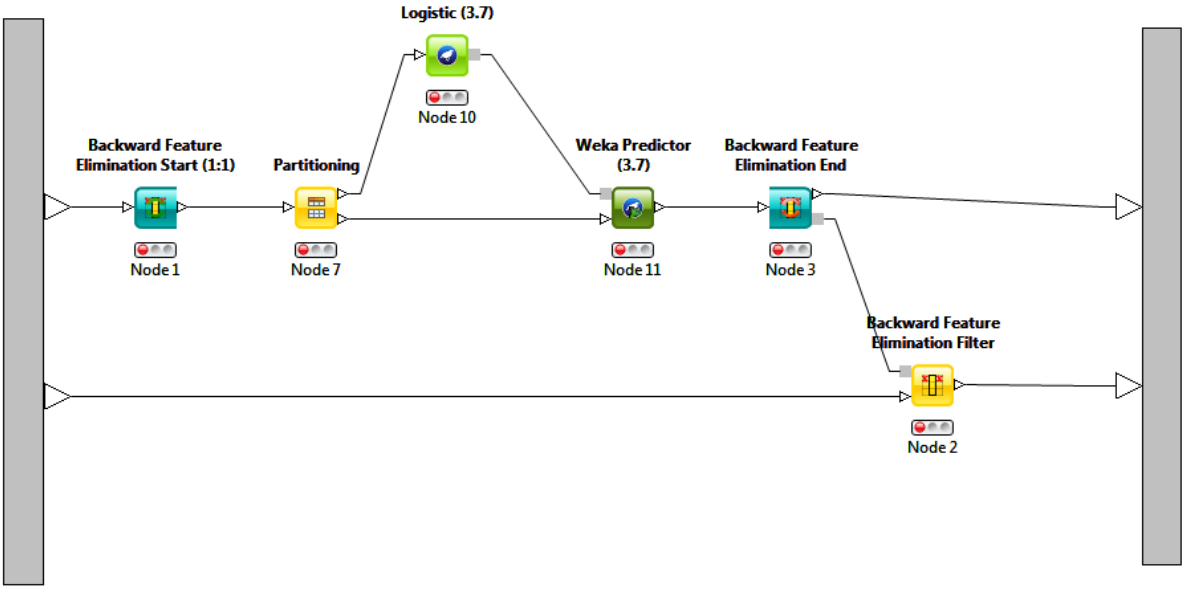

Figure A. 2: KNIME workflow implementing the backward feature selection for the Logistic regression classifier

Table A. 2: Descriptors selected by backward feature selection for A-B classification with Logistic regression

Molecular Descriptor
Aliphatic ring count
Acceptor count
Strongest BASIC pKa 2
logD at pH=5.0
logD at pH=7.4
Polar surface area
#atoms in 4 ring
#C = bonded to C and 3 heavy atoms
#atoms in 3 ring
#C bonded to at least 3 N atoms
#QH 3 bonds from another QH
#non-ring bonds that connect rings
#S in double/charge separated bonds
#N non-ring bonded to a ring
#C in C=C bonded to >= 3 heavy atoms
#N separated by 4 bonds
#heteroatoms in 5 ring
#XQ>3 bonded to at least 1 halogen
#N in double bonds
#het-het bonds
#CH2s separated by 4 bonds
#halogens

Total # ring HETEROCYCLE atoms
#OH groups
(key(160)-1 if key(160)>1; else 0) Key160 = #CH3 groups
#O in C=O
#XN where coord. # of X>=3
#N in C-N single bonds
Key(164)-1 if key(164)>1; else 0 key164 = #Oxygens
#N
#ring atoms

Table A. 3: Descriptors selected by backward feature selection for B-A classification with Logistic regression

Molecular Descriptor
Fused aliphatic ring count
Strongest ACIDIC pKa 2
Strongest BASIC pKa 2
logD at pH=6.5
#C bonded to at least 3 N atoms
#S atoms bonded to N
#N in C#N
#O in rings
#N non-ring bonded to a ring
#CH2 or CH3 separated by non-C
#halogens bonded to any ring
#methylated heteroatoms
#atoms in 5-rings
#N attached to CH2
#O separated by 1 C
#CH2s separated by 4 bonds
#CH2s separated by 3 bonds
(# het atoms with H)
#N non-ring bonded to a ring
Bit: is there more than 1 O=
Total # ring HETEROCYCLE atoms
#O in C=O
#non-ring CH2
#O in C-O single bonds
#N in C-N single bonds
#N
#ring atoms

Table A. 4: Descriptors selected by the backward feature selection for efflux ratio classification

Descriptors
#N
#halogens
#CH2s separated by 3 bonds
#heteroatoms in 5ring
#non-C with coordination number >=3
#O in C=O
#N attached to CH2
#CH2s separated by 4 bonds
#N bonded to >= 3 C
#methylated heteroatoms
Bit: is there more than 1 O=
#S in double/charge separated bonds
#CH3 groups
#N attached to CH2
#atoms in 5 membered rings
#atoms separated by (!:):(!:)
Strongest ACIDIC pKa 1
#OH groups
#C in C=C
#heteroatoms in 5 ring