Development of a correlation based and a decision tree based prediction algorithm for tissue to plasma partition coefficients

by

Yejin Yun

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Master of Science

in

Pharmacy

Waterloo, Ontario, Canada, 2013

AUTHOR'S DECLARATION

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Physiologically based pharmacokinetic (PBPK) modeling is a tool used in drug discovery and human health risk assessment. PBPK models are mathematical representations of the anatomy, physiology and biochemistry of an organism. PBPK models, using both compound and physiologic inputs, are used to predict a drug's pharmacokinetics in various situations. Tissue to plasma partition coefficients (K_p), a key PBPK model input, define the steady state concentration differential between the tissue and plasma and are used to predict the volume of distribution. Experimental determination of these parameters once limited the development of PBPK models however *in silico* prediction methods were introduced to overcome this issue. The developed algorithms vary in input parameters and prediction accuracy and none are considered standard, warranting further research. Chapter 2 presents a newly developed K_p prediction algorithm that requires only readily available input parameters. Using a test dataset, this K_p prediction algorithm demonstrated good prediction accuracy and greater prediction accuracy than preexisting algorithms. Chapter 3 introduced a decision tree based K_p prediction method. In this novel approach, six previously published algorithms, including the one developed in Chapter 2, were utilized. The aim of the developed classifier was to identify the most accurate tissue-specific K_p prediction algorithm for a new drug. A dataset consisting of 122 drugs was used to train the classifier and identify the most accurate K_p prediction algorithm for a certain physico-chemical space. Three versions of tissue specific classifiers were developed and were dependent on the necessary inputs. The use of the classifier resulted in a better prediction accuracy as compared to the use of any single K_p prediction algorithm for all tissues; the current mode of use in PBPK

model building. With built-in estimation equations for those input parameters not necessarily available, this K_p prediction tool will provide K_p prediction when only limited input parameters are available. The two presented innovative methods will improve tissue distribution prediction accuracy thus enhancing the confidence in PBPK modeling outputs.

Acknowledgements

I would like to express my gratitude to Dr. Andrea Edginton for her support as my thesis supervisor. I would like to thank my committee members Dr. Ceclilia Cotton and Dr. Shawn Wettig for their advice throughout the project. I would like to thank my colleagues for their moral support.

I thank Dr. Walter Schmitt, Dr. Ramus Jansson and Dr. Kannan Krishnan for providing their data files.

Dedication

I dedicate my work to my family and many friends.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AAFE	Absolute average fold error
ADME	Absorption, distribution, metabolism and elimination
AFE	Average fold error
AIC	Akaike information criterion
B:P	Blood to plasma ratio
Bagging	Bootstrap aggregation
BBB	Blood brain barrier
E	Extraction ratio
FE	Fold error
Fi	Fraction of ionized drug
fup	Unbound fraction in plasma
HSA	Human albumin serum
K_p	Tissue-to-plasma partition coefficient
K_pu	Tissue-to-plasma water partition coefficient
K_puBC	Unbound compound concentration in blood cells
LI	Large intestine
LogD	Logarithmic value of n-octanol-water partition coefficient adjusted for ionization at pH 7.4
LogKvo:w	Logarithmic value of vegetable oil-water partitioning adjusted for ionization at pH 7.4

LogP	Logarithmic value of N-octanol-water partition coefficient
M	The number of variables
MA	Membrane affinity
MFE	Mean fold error
m_{try}	Optimal value of the number of variables
n_{tree}	Number of trees
Obs	Observed K_p values
OOB	Out of bag
PBPK	Physiologically based pharmacokinetic
PC	Partition coefficient
Pgp	P-glycoprotein
PhS	Phosphophatidyl serine
PK	Pharmacokinetics
Pred	Predicted K_p values
R^2	Coefficient of determination
RBCu	Red blood cell partitioning data for unbound drugs
RMSE	Root mean square error
SI	Small intestine
SPR	Surface plasmon resonance
TCB	Tissue composition based
VIF	Variance inflation factor
Vss	Volume of distribution at steady state

Chapter 1

Introduction

Physiologically-based pharmacokinetic (PBPK) modeling

Pharmacokinetics (PK) is the mathematical description of the absorption, distribution, metabolism and excretion (ADME) of a compound and a quantitative description of how these processes affect the time course and intensity of response. One means of predicting and assessing the pharmacokinetics of a compound is through the use of PBPK modeling. As a result, PBPK models are used in pharmaceutical research, drug development and in toxicological risk assessment. PBPK models are mathematical constructions that are developed to represent the organism of interest. A whole body PBPK model is comprised of physiological compartments that represent organs or tissues (Figure 1-1). Each organ is represented as either one well-stirred compartment (e.g. one homogenously mixed unit) or as multiple compartments that represent, for example, vascular, interstitial and/or intracellular space. Organ compartments are linked together through venous and arterial blood pools with closure of the system through the lungs. Mass transfer between each compartment identified in the model is represented using a differential equation such that the entire PBPK model becomes a series of differential equations.

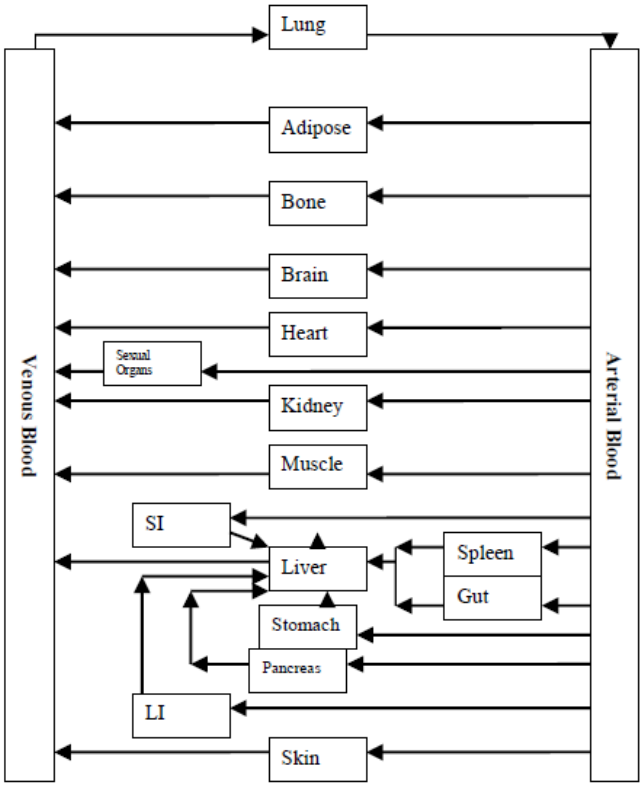


Figure 1-1. Structure of PBPK model. (SI: Small intestine, LI: large intestine)

Each organ compartment within the PBPK model is defined by a species specific blood flow rate (the sum of which equals the total cardiac output) and a physiologic volume [1]. Compound specific parameters such as protein binding affinity, tissue to plasma partition coefficients, clearance and permeability x surface area products (if organs are not considered well-stirred) are required for the initial parameterization of a PBPK model. Once a PBPK model is structured and parameterized, simulations under various dosing regimens or conditions can be made.

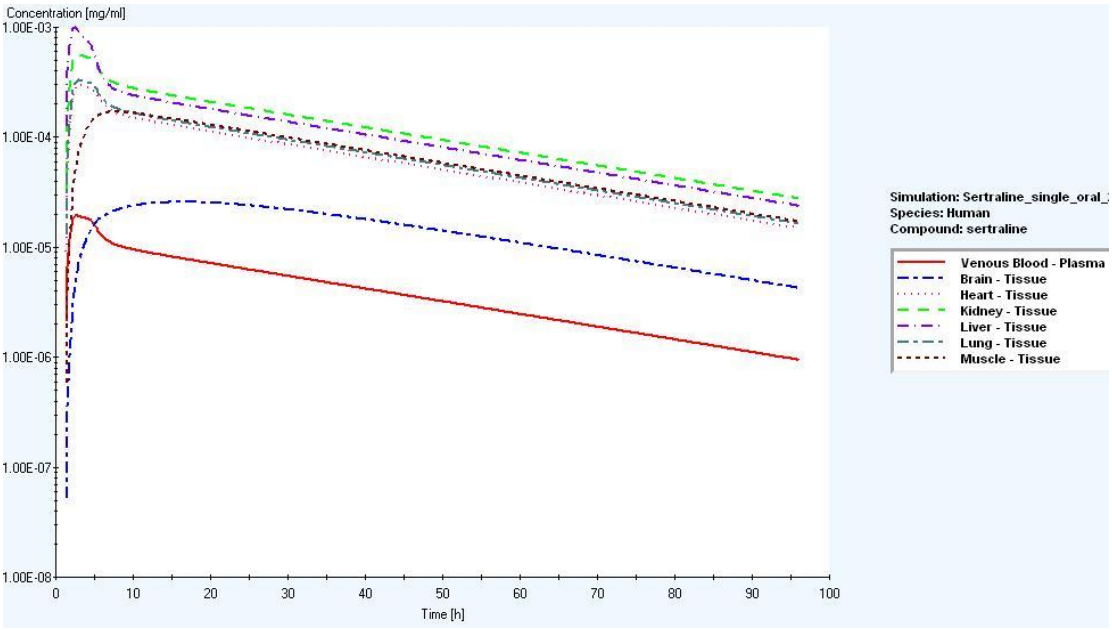


Figure 1-2. An example of simulated concentration versus time profile in tissues and the plasma by a PBPK model

In early drug discovery, a drug candidate is screened among thousands of possible compounds. The empirical approach in the selection of a drug candidate can be time consuming, labor intensive and costly. Therefore, drug candidate screening and a first-time-in-animal study design can be aided by PBPK modeling for the prediction and understanding of a compound's ADME. Furthermore, PBPK models predict the human PK as early as possible which can help to identify undesirable PK characteristics of a drug candidate. Early PK prediction can help to reduce the cost associated with drug development and potentially reduce the rate of failure in drug development.

The following steps are taken in PBPK modeling for interspecies scaling. Once compound specific parameters (e.g. unbound fraction in plasma, species specific clearance) and species specific anatomical and physiological parameters are input, a series of concentration vs. time profiles are simulated for any organ or tissue that is included in the model (e.g. Figure 1-2). To

ensure appropriate distribution and clearance, a comparison of the simulated and the experimentally determined profiles are made. Uncertain input parameters are optimized (e.g. tissue to plasma partition coefficients) until there is adequate agreement between the simulated and experimentally determined curves. This usually occurs in the rat. Scaling to humans is then completed by replacing the anatomical, physiological and biochemical inputs to that of humans and re-simulating. This provides a biologically rational approach to interspecies scaling of PK.

Tissue distribution

The distribution of a compound within a system (i.e. tissue distribution) is the process of compound partitioning into the tissues from the systemic circulation. Compound properties (e.g. lipophilicity) and the nature of tissue cellular membranes determine the ability of the compound to permeate into the tissue. For example, lipophilic compounds tend to partition to a greater extent into lipid-rich tissues such as adipose and brain whereas hydrophilic compounds tend to distribute into lean tissues such as heart and muscle. The extent of tissue distribution is dependent on tissue partitioning and the binding affinity of a compound to blood cells, proteins and tissue components [1]. The global parameter that quantifies the extent of compound distribution from plasma into tissues is the volume of distribution at steady state (Vss). This is a PBPK modeling output. For example, a small Vss indicates a lack of tissue specific binding and/or an affinity for binding to plasma proteins. Compounds with a large Vss have extensive affinity for binding in tissues.

Due to various tissue compositions, compound concentration is tissue-specific. The extent of compound distribution into an individual tissue is expressed by a steady state tissue to plasma

partition coefficient (K_p), i.e. the ratio of the concentration of a compound in tissue and plasma [2]. Thus, the relationship between Vss and K_p is expressed as Eqn.1-1 [3]:

Eqn. 1-1 $$Vss = V_{plasma} + \sum_{1}^{n} Kp_i \times V_{tissue,i} \times (1 - E_i)$$

where V_{plasma} and V_{tissue}, is the physiologic volume of plasma and respective tissue. E is the extraction ratio of an eliminating tissue (i.e. the liver or the kidneys) and is a measure that represents the ability of a tissue to remove a compound from the systemic circulation through excretion in the urine or enzymatic metabolism in the liver. For non-eliminating tissue, extraction ratio is zero (E_i=0).

K_ps are used to quantify the extent of a compounds distribution from the systemic circulation into the tissues at steady-state. The K_ps used in PBPK models comprise the tissue: plasma partition coefficients based on total (K_p) [2,4-6] or unbound concentration (K_pu) [7-9] in the case of drug compounds or the tissue: blood partition coefficients [10] based on total concentration for environmental chemicals. The tissue distribution prediction within a PBPK model is sensitive to the K_p values. Historically, these values were derived experimentally *in vivo*. This is a costly and time consuming endeavor and has been a limitation in the development of PBPK models. As a result, K_p prediction algorithms using *in vitro* and *in silico* data have been developed to overcome the need for experimental K_p determination. These algorithms predict K_ps based on the underlying physiology and behavior of a compound in the body.

K_p prediction algorithms are divided into two areas: (i) tissue composition based (TCB) algorithms that are created solely using physico-chemical properties of the compound along with tissue specific parameters and (ii) correlation based algorithms that are empirically derived using both compound specific information and information derived *in vivo* (e.g. muscle K_p).

Tissue composition based algorithms

TCB algorithms are mechanistic in nature and do not require *in vivo* information as input. In early studies, tissue solubility of a compound was calculated by assuming: (i) solubility of a chemical in n-octanol corresponds to its solubility in tissue neutral lipids, (ii) solubility in water corresponds to water fraction and (iii) solubility in phospholipids is a function of solubility in water and n-octanol [10]. Using this assumption, the solubility of a chemical in tissue was then calculated as the sum of the solubilities listed above [11]. Building on this, a mechanistic model based on tissue composition, physico-chemistry, and plasma protein binding was developed by Poulin and his coworkers and later revised by Berezhkovskiy [12]. The main assumption of this TCB model is that the distribution of a compound is primarily governed by passive diffusion into tissue compartments and reversible binding to common proteins that are in the plasma and tissue interstitial spaces.

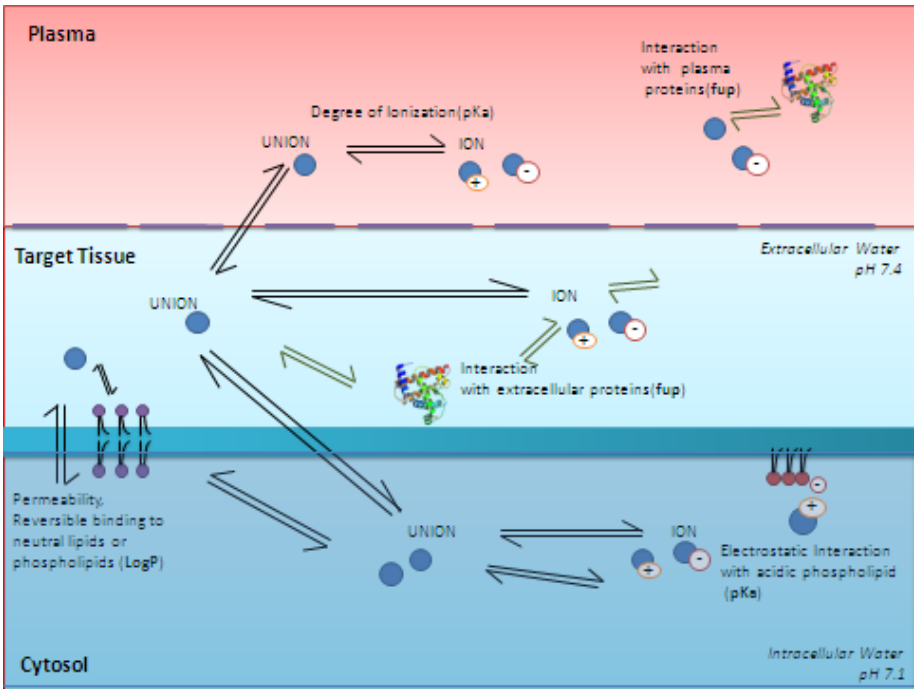


Figure 1-3. A schematic showing the underlying processes of tissue partitioning that were described by Rodgers *et al.* model [8,9].

Later, Rodgers and Rowland (2005a) extended and enhanced the TCB model by incorporating the electrostatic interactions of moderate to strong bases (pKa ≥ 7) with acidic phospholipids to predict K_pu. This model assumes that the electrostatic interactions prevail and compounds distribute passively into intra- and/or extracellular tissue water. The equation also accounts for two processes: (i) dissolution of both ionized and unionized portions of a compound into tissue water and (ii) partitioning of unionized compounds into neutral lipids and neutral phospholipids (Figure 1-3). The researchers also attempted to predict K_pu which is the steady state parameter that relates the unbound concentration in tissues to unbound concentration in plasma. The reason for predicting K_pu as opposed to K_p is that only unbound compounds can distribute across cellular membranes.

Rodgers and coworker(s) [9] continued to develop a new mechanistic equation for predicting the K_ps for neutrals, acids, and weak bases by considering the compound interactions with proteins. This is an important factor for the tissue distribution of compounds because of the abundance of proteins that are present in the extracellular space. Lipophilic neutrals preferentially bind to lipoproteins, whereas acids and weak bases primarily bind to albumin. Zwitterions can be divided into two groups. The first group includes compounds with one basic form (pKa ≥ 7), thus it is presumed to undergo interactions with acidic phospholipids in the same manner that strong bases do. The second group consists of all other zwitter-ionic compounds and they are thought to have the same distributional behavior as acids and very weak bases [8,9]. Therefore, the degree of the affinity of the compounds to the extracellular proteins is a crucial parameter in the prediction of K_ps.

Schmitt [6] built a TCB algorithm to calculate K_ps of classes of compounds based on their lipophilicity, pKa, binding ability to phospholipids and the unbound fraction in plasma. Specifically, compound binding to phospholipids was explained in a mechanistic way by accounting for the interaction between charged phospholipids and charged molecules along with consideration of the phosphatidylcholine:buffer partition coefficient and the phospholipid:water partition coefficient. This model can be applied universally for all classes of compounds, which implies the significance of this algorithm. Later, Peryet and his coworkers [13] developed the algorithm that unifies the mechanisms involved in the distribution of both drug compounds and environmental chemicals. The unified algorithm provides predictions of K_ps by calculating the ratio of the concentration in cellular and interstitial space to the concentration in plasma and red blood cells (RBC). The Peryet et al. (2010) algorithm also accounted for the consideration of different volumes in each matrix. The researchers attempted to integrate and reproduce the previously published equations into a single algorithm. Their calculations yielded the same level of accuracy when compared to previous studies. In addition, this unified algorithm predicts partition coefficients at both the macro (tissue: plasma partition coefficient) and the micro (cells: fluid partition coefficient) levels [13].

Correlation based algorithms

The relationship between experimentally determined *in vivo* parameters (e.g. a muscle K_p) and K_ps has been utilized to develop predictive regression equations to estimate K_ps. The work of Bjorkman [4] demonstrated that muscle K_p can be used to represent other tissue K_ps. Specifically, lean tissue K_ps can be calculated using a linear regression equation with muscle K_p as a predictor. The empirical method was later refined by the work of Jansson [5]. For this model, the

relationship between muscle K_p and non-adipose K_p was improved by incorporating compound lipophilicity data into the equations.

For moderate-to-strong bases, it was observed that the K_p predictions were less accurate than for neutral, acidic and weakly basic compounds [8]. This was mainly due to their ionic interaction with acidic phospholipids such as phosphophatidyl serine (PhS). The work of Yata and colleagues demonstrated that the inter-organ variation in tissue distribution of basic compounds varies with PhS concentration [14]. The study of Poulin and Theil introduced a correlation based algorithm that utilized red blood cell partitioning data for unbound compounds (RBCu) [7]. RBCu was determined *in vitro* and used as an indicator of the degree of binding capacity due to electronic interactions of basic compounds with acidic PhS. The rationale for this correlation is that RBCs are rich in acidic phospholipids and the membrane of RBCs play a similar role to the cellular membrane in a lean tissue. In this study, the relationship between RBCu and tissue K_ps as well as the relationship between muscle K_ps and tissue K_ps was used to develop predictive regression equations. It was observed that K_p prediction with muscle K_p as a predictor was more accurate than the use of RBCu as a predictor alone [7]. This approach was further enhanced by identifying outliers of the over-prediction of K_ps. Both pharmacological activity of a compound and compound specific properties such as pKa and lipophilicity were taken into account to refine the correlation approach of the Poulin and Theil model [7,15].

Input parameters for K_p algorithms

Various input parameters for the introduced algorithms are often determined *in vitro* and used in TCB algorithms to estimate: (i) the hydrophobic interactions of a compound with neutral phospholipids (e.g. n-octanol: buffer partition coefficient, or vegetable oil: buffer partition

coefficient), (ii) the ionic interaction with charged phospholipids, (iii) hydrophobic binding to hemoglobin (e.g. blood: air partition coefficient) and (iv) the binding to plasma proteins (e.g. unbound fraction in plasma). Some of the important parameters in the previously explained algorithms are described below.

Lipophilicity is one of the most important ADME-related properties and has a major impact on pharmacokinetics. Lipophilicity of a compound is determined using LogP from octanol/water partitioning. LogP is the logarithm of the partition coefficient of the compound trapped between an organic phase and an aqueous phase at a pH where all of the compounds are in their neutral forms. N-octanol is thought to mimic the hydro-lipophilicity balance of neutral lipid mixtures; therefore, the distribution of a compound into n-octanol was postulated to simulate the ability of a compound to passively diffuse across biological membranes. However, n-octanol is not a suitable surrogate to mimic the triglycerides of adipose tissue. The solution to this would be to use olive oil, which is abundant in triglycerides. Therefore the logarithm of olive oil: buffer partition coefficient (LogKvo:w) provides a more accurate K_p prediction for adipose tissue [8,9,16]. Additionally, LogD is the logarithm of the distribution coefficient of the compound at a specific pH. LogD depends on the partitioning of the ionized portion of the molecules and the partitioning of the neutral portion of the molecules.

The fraction of unbound compound in plasma (fup) is also an important descriptor in K_p prediction models. Binding of a compound to plasma proteins affects its distribution. The degree of binding is frequently expressed as a ratio of bound to total concentration. The unbound fraction of a compound is the proportion of the compound in plasma or in tissue interstitial space that is not bound to common proteins such as albumin, glycoproteins, lipoproteins and globulins. The steady-state concentration of an unbound compound is equal in all body tissues, regardless

to the degree of the binding to the macromolecules. Therefore, the value of K_p can be defined as the ratio of the fraction of unbound compound in the plasma to the fraction of unbound compound in the tissue. Furthermore, the fraction of the unbound compound is regarded to be pharmacologically active. Since bound protein-compound complexes cannot penetrate the capillary membrane, the rate of distribution of compound into tissue is dependent on the concentration gradient produced by the concentration of unbound unionized compound.

A molecule's pKa, which is a determining factor in the degree of ionization at a particular pH, is a key chemical property in K_p predictions. Compounds that are weak acids or weak bases exist in solution at equilibrium between the unionized and ionized form. Only un-ionized nonpolar chemicals can cross the tissue membrane as ionized compounds are less permeable than un-ionized compounds. At equilibrium, the concentrations of the un-ionized compounds are equal in both plasma and tissue. However, total concentration in one matrix (e.g. a tissue) may be different depending on the degree of ionization of a compound at a tissue-specific physiological pH. For the statistical analyses in this study, a variable that indicates the degree of ionization of a compound as a function of tissue pH is needed. The ionized fraction of the compound (fi) represents the degree of ionization at a tissue-specific physiological pH (equations are presented in the chapter 2). The fi equations are derived from the Henderson-Hasselbalch equation. The fi value ranges from 0 to 1 where a highly ionized compound at specific pH approaches 1.

Figure 1‑4 presents the simulation of fi value at various compound pKas. The influence of different tissue pH is demonstrated (i.e. pH 7.4 for plasma, pH 6.6 for lung). For a compound with an acidic pKa where the pKa value is smaller than the tissue pH, the fi is high (Figure 1‑4, top). For a compound with a basic pKa where the pKa value is larger than the tissue pH, the fi is high (Figure 1‑4, bottom). With knowledge of pKa (acidic or basic pKa), this variable can

distinguish the ionized fraction for compounds with the same value of pKa. For example, for a compound with acidic pKa of 7, the fi value at the plasma pH 7.4 is 0.72. For a compound with basic pKa of 7, the fi value at the plasma pH 7.4 is 0.28. In addition, for a neutral compound, the fi value is zero. Thus, fi is considered to be a better representative parameter for describing a compound's degree of ionization at various tissue pH than the use of pKa alone.

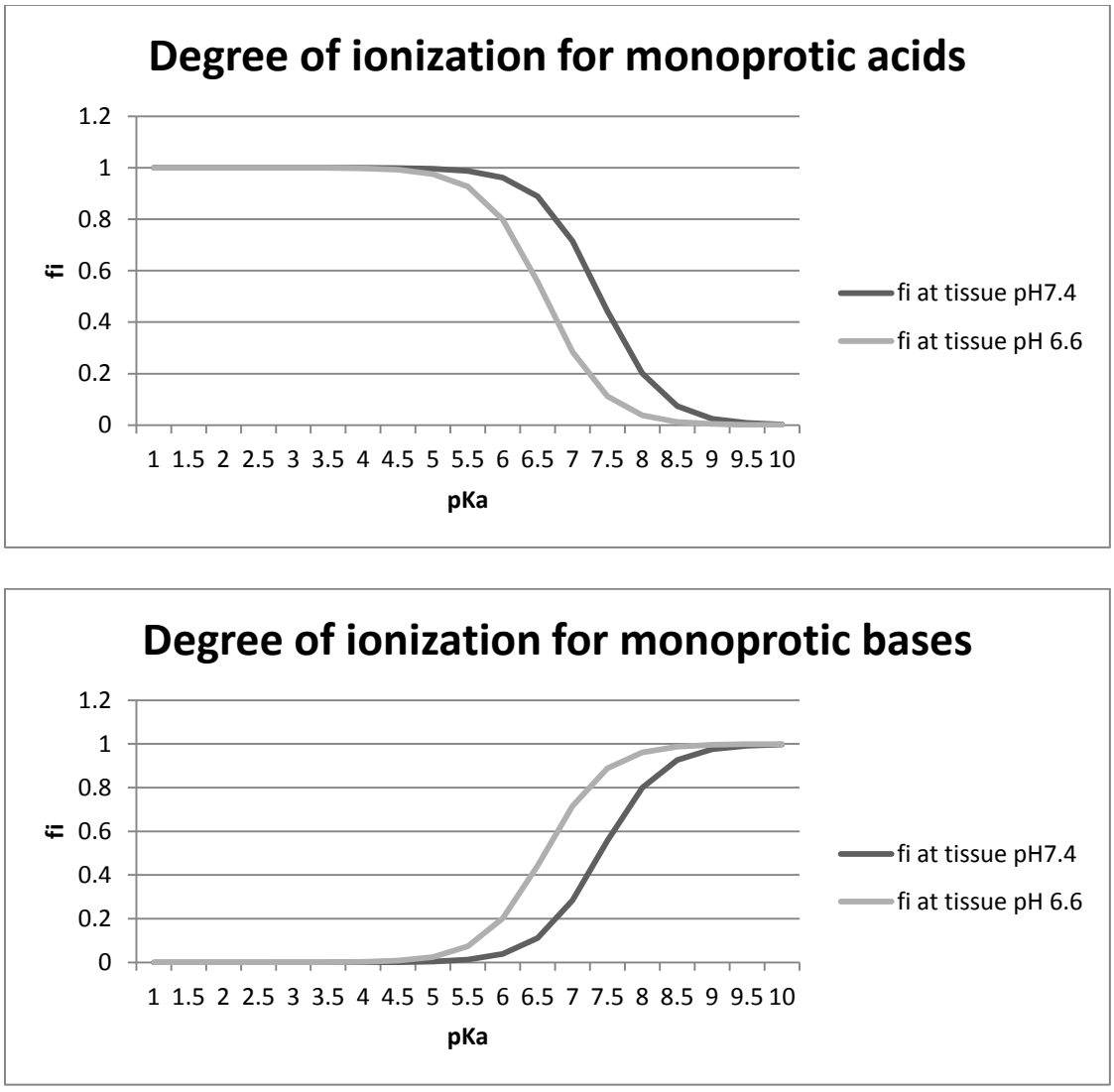


Figure 1-4. Simulation of degree of ionization at various tissue pH for monoprotic acids (top) and monoprotic bases (bottom).

Compound affinity to red blood cells is often used as an indicator of *in vivo* distribution. It has been observed that a compound's ability to bind to hemoglobin within RBCs correlates with the lipophilicity of the compound [17]. Compound binding to RBCs is a crucial factor in representing tissue distribution because RBCs are rich in acidic phospholipids, which are responsible for the high binding affinity of basic compounds. Only a few algorithms (e.g. [7,8]) require RBCu. Poulin and Theil [7] demonstrated that the K_p prediction with muscle K_p as an input variable was more accurate than the K_p prediction with RBCu as an input variable. The muscle K_p is also an important factor in K_p prediction since muscle is a highly perfused organ, and accounts for approximately 40% of the total body mass. For compounds with a large Vss, a substantial portion of the compound is considered to partition into the muscle. In addition, Vss also can be used as an input as it is the parameter that represents the overall extent of the drug distribution in the body [5,18,19].

These physico-chemical and physiological inputs represent key input parameters for K_p prediction algorithms. Some of these input parameters are readily available such as a measure of lipophilicity or pKa while others are not routinely measured such as RBCu or muscle K_p. Due to the difficulty in obtaining some of the input parameters; several algorithms have limited utility in tissue-specific K_p prediction for a novel compound.

Thesis objectives

This thesis aims to enhance the confidence in K_p predictions. First, a novel correlation based prediction algorithm is developed that uses readily available inputs. The hypothesis for this study was that this correlation based algorithm will increase the tissue specific accuracy in K_p prediction for a tissue.

Second, a machine learning method is used to develop a decision tree that will select, for each tissue, the best-predicting K_p algorithm. This will allow the user to harness the best of all of the algorithms for their novel compound. The hypothesis for this study was that *the use of a decision tree will produce a more accurate overall prediction of K_ps than any one K_p prediction algorithm alone.* This will result in an adequate parameterization of a PBPK model. These two innovative methods will improve tissue distribution prediction accuracy therefore enhancing the confidence in PBPK modeling outputs.

Chapter 2

Correlation-based prediction of tissue-to-plasma partition coefficients using readily available input parameters[a]

2.1 Outline

1. Rationale: Tissue-to-plasma partition coefficients (K_p) that characterize the tissue distribution of a drug are important input parameters in physiologically based pharmacokinetic (PBPK) models. The aim of this study was to develop an empirically derived K_p prediction algorithm using input parameters that are available early in the investigation of a compound.

2. Methods: The algorithm development dataset (n = 97 compounds) was divided according to acidic/basic properties. Using multiple stepwise regression, the experimentally derived K_p values were correlated with the rat volume of distribution at steady state (V_{ss}) and one or more physicochemical parameters (e.g., lipophilicity, degree of ionization, protein binding) to account for inter-organ variability of tissue distribution.

3. Results: Prediction equations for the value of K_p were developed for 11 tissues. Validation of this model using a test dataset (n = 20 compounds) demonstrated that 65% of the predicted K_p values were within a two-fold error deviation from the experimental values. The developed algorithms had greater prediction accuracy compared to an existing empirically derived and a mechanistic tissue-composition algorithm.

[a] Chapter 2 has been published in the journal Xenobiotica.
Yun, Y. E. & Edginton, A. N. 2013, "Correlation-based prediction of tissue-to-plasma partition coefficients using readily available input parameters", Xenobiotica 43: (In press). doi 10.3109/00498254.2013.770182

4. Conclusions: This innovative method uses readily available input parameters with reasonable prediction accuracy and will thus enhance both the usability and the confidence in the outputs of PBPK models.

2.2 Introduction

Physiologically based pharmacokinetic (PBPK) modeling is widely used in pharmaceutical research, drug development and toxicological risk assessment to make predictions of the target tissue exposure following various administration scenarios [20]. An inherent advantage of PBPK approaches is the ability to incorporate both intrinsic (e.g., age, organ dysfunction [21,22]) and extrinsic (e.g., drug-drug interaction [23]) factors into the models, which provides the ability to make biologically plausible PK predictions and extrapolations across and within species [24]. A PBPK model uses anatomically and physiologically appropriate compartments of a body (e.g., tissues), which are linked through systemic circulation with the system closed through the lung [3,25-27]. In addition to organ-specific inputs, PBPK models also require drug-specific inputs, such as a measure of the binding affinity to plasma proteins (fu_p), the tissue to plasma partition coefficients (K_p), the permeability × surface area products and the drug dissolution properties. Although anthropometric parameters are available for many organisms, drug-specific inputs are more uncertain and just as crucial to the success of the model prediction.

One of the most important drug-specific input parameters is the tissue to plasma partition coefficients, K_p, i.e., the ratio of the concentration of a compound in the tissue to the concentration of the compound in the plasma at steady state [2,6]. The value of this coefficient indicates the degree of accumulation of a drug in a tissue under steady-state conditions and

represents the relative exposure of a drug between different tissues, which enables a target site-related assessment of absorption, distribution and elimination [28].

The K_p values partially define the volume of the distribution at steady state (V_{ss}), which is the ratio of the total amount of the drug in the body to the total amount of the drug in the plasma under steady-state conditions [29-31]. The V_{ss} value represents the overall extent of the drug distribution in the body and is defined as in Eqn. 2-1:

Eqn. 2-1 $$Vss = V_{plasma} + \sum_{1}^{n} Kp_i \times V_{tissue,i} \times (1 - E_i)$$

where V_{plasma} is the volume of the plasma and $V_{tissue,i}$ is the volume of the i^{th} tissue. For non-eliminating tissues, extraction ratio E_i is zero (E_i=0). If the model is parameterized with the appropriate K_p values, PBPK models can predict the V_{ss} because the plasma and tissue volumes are inherent parameters in the model.

The K_p values can be experimentally derived in rodents through destructive sampling and are generally considered to be the most desirable input parameters because their uncertainty is low. However, the experimental *in vivo* determination of these K_p values can be misleading if steady state is not reached at the time of the measurement; for example, highly lipophilic molecules require a longer time to reach steady state than the time that researchers might be willing to wait. This experimental determination of these parameters is also time consuming and expensive [6,32]. As a result, to minimize the need of experimental procedures in animals, algorithms that predict the K_p values based on the physico-chemical characteristics of the compound and organism-specific parameters have been developed. Two types of algorithms exist: mechanistic algorithms and empirically derived algorithms.

Tissue composition-based (TCB) algorithms are mechanistic in nature and provide initial estimation of the K_p values when *in vivo* information (e.g., muscle K_p) is unavailable. TCB modeling aims to describe the combination of the interactions that occur in any one tissue as a result of the physiological components of the tissue and the chemical properties of the compound [2,6,8,9]. In early TCB models [2,2,10], the tissue-to-blood partition coefficients were predicted by estimating the ratio of the solubility of a chemical in tissues to that in blood. The solubility in each matrix was approximated as the total solubility of the compound in neutral lipids, phospholipids, and water. Rodgers et al. enhanced these models by incorporating the electrostatic interactions of basic compounds (pKa $\geq$ 7) with cellular acidic phospholipids [8]. With neutral, acidic, and weak basic compounds, the prediction of K_p values is primarily defined by their interaction with extracellular proteins (i.e., lipoproteins, albumin) [9]. Further modifications to the model were made by Schmitt [6], who accounted for the combination of the effects of the drug distribution in the interstitial space, the effects of the pH gradient between the plasma and the tissues, the partitioning into the different lipid components in the tissues, and the binding to proteins.

Correlation-based K_p prediction models are empirical in nature and offer an alternative approach to the TCB models. These correlation-based models use both physicochemical descriptors of a compound [5] and organism-specific data, such as muscle K_p [4,5,7] and red blood cell partitioning data [7] as predictor variables. Early correlation-based models used an experimentally determined muscle K_p value that was correlated with other tissue K_p values through regression [29]. Bjorkman [4] performed similar work but also used adipose K_p values as a predictor [4]. The work of Jansson et al. [5] enhanced Poulin and Theil's [29] approach by incorporating the compound lipophilicity (i.e., LogP, $LogD_{7.4}$ or $LogK_{7.4}$) as a secondary predictor. Jansson et al.

[5] uses $K_{p,muscle}$ as the ultimate input parameter. If the value of $K_{p,muscle}$ is not available, equation 2 can be used to generate the value of this parameter from V_{ss}.

Eqn. 2-2 $$Vss = V_{plasma} + \sum_{1}^{n} V_{tissue,i} \times 10^{a \times \log(Kp,muscle) + b \times \log(lipophiliaty) + c}$$ [5]

The Jansson et al. [5] method requires either an experimentally derived value for the muscle K_p or the value of V_{ss}, which can be used to predict the value of $K_{p,muscle}$. This parameter is then used in the regression equations. Poulin and Theil [7] proposed a correlation model that utilized red blood cell partitioning data for unbound drugs (RBCu) as an indicator of the degree of the binding capacity of basic drugs with acidic phosphatidylserines.

Recently, a comparison of the current methods for the determination of V_{ss} based on the estimation of K_p and the use of Eqn.2-2 found that the correlation-based models, especially Jansson et al. model [5], were more accurate than even the best TCB model, which was developed by Rodgers *et al*. [33,34]. The results suggest that the correlation-based methods have a higher accuracy in K_p prediction; however, these models also require input parameters (i.e., muscle K_p and RBC_u) that are difficult to obtain and not regularly measured. The V_{ss} in rats is a readily available parameter; therefore, PK studies in rats are completed relatively early in the drug discovery process and are commonly completed for environmental xenobiotics [18]. The current study aims to develop a correlation-based K_p prediction model that directly uses the rat V_{ss} as a primary K_p predictor and links this value with secondary physicochemical parameters for tissue-specific K_p estimation.

2.3 Methods

Drug specific parameters

The drug-specific parameters that affect the tissue distribution are the lipophilicity, the degree of ionization, and the plasma protein binding. In this model, the distribution of a drug into and out of a tissue was solely attributed to passive diffusion.

The lipophilicity of a drug, which is one of the most important ADME-related properties, has a major effect on its pharmacokinetics. The lipophilic or hydrophilic properties of a drug can be described by the N-octanol-water partition coefficient (LogP). N-octanol is considered to imitate the hydro-lipophilicity balance of biological membranes because it contains a saturated alkyl chain and a hydroxyl group and has a similar solubility in water [29,35]. In general, a high lipid solubility leads to a high affinity to neutral lipids, proteins and other macromolecules, which ultimately imparts extensive drug distribution [36]. LogP values were incorporated into the statistical analysis to account for a drug's affinity to the lipophilic constituents of a tissue.

Table 2-1. Tissue pH values in rats

Tissue	pH[a]
Adipose	7.1
Bone	7
Brain	7.1
Gut	7
Heart	7.1
Kidneys	7.22
Liver	7.1
Lung	6.6
Muscle	6.81
Skin	7
Spleen	7

[a]Obtained from the literature[6,37-43].

The tissue distribution is greatly affected by the acidic/basic properties of the compound. It is hypothesized that an electrostatic interaction between the cellular acidic phosphatidylserine and the basic moiety of a drug is crucial to the definition of the tissue distribution of moderately to strongly basic drugs [7,8,44]. However, acidic, weakly basic and neutral compounds are known to bind to extracellular proteins: acids and weak bases bind to albumin and lipophilic neutrals bind to lipoproteins (Rodgers & Rowland 2006). These classes of compounds tend to have smaller distribution volumes than moderate to strong bases [9,45]. As a result, compounds were considered in two groups: moderate to strong bases and acidic, neutral and weak bases (see below).

The degree of ionization is an important factor in tissue distribution. This is mainly due to the differential pH between the plasma/interstitial space and the intracellular water space. As shown in table 2-1, the pH of tissues is lower than the plasma pH (7.4) and varies across the tissue. Therefore, the influence of the degree of ionization on the distribution is different for each tissue. To account for the inter-tissue distribution variation, the ionized fraction of the drug (fi) was calculated (Eqn 2-3 to 2-7); these values represent the degree of ionization at a tissue-specific physiological pH:

Eqn. 2-3 $fi = 1 - [1 + 10^{pKa - pH\ tissue}]^{-1}$ for monoprotic bases,

Eqn. 2-4 $fi = 1 - [10^{pKa_1 - pH\ tissue} + 10^{pKa_1 + pKa_2 - pH\ tissue \times 2}]^{-1}$ for diprotic bases,

Eqn. 2-5 $fi = 1 - [10^{pH\ tissue - pKa}]^{-1}$ for monoprotic acids,

Eqn. 2-6 $fi = 1 - [10^{pH\ tissue - pKa_1} + 10^{pH\ tissue \times 2 - pKa_1 - pKa_2}]^{-1}$ for diprotic acids,

Eqn. 2-7 $fi = 1 - [10^{pka_{base} - pH\ tissue} + 10^{pH\ tissue - pKa_{acid}}]^{-1}$ for zwitterions.

The tissue-specific fi values of each compound were incorporated into the statistical analysis as potential predictor variables.

The steady-state concentration of an unbound drug is equal in all of the body tissues, regardless of the degree of the binding to macromolecules [46]. Therefore, the value of K_p can be defined as the ratio of the fraction of unbound drug in the plasma to the fraction of unbound drug in the tissue. The unbound fraction in the plasma (fu_p) was therefore incorporated into the statistical analysis as a potential predictor of K_p.

Data collection

A database of the experimentally derived K_p values, the rat V_{ss} and the corresponding physicochemical properties was created from the literature (Appendix 1-4). Additional criteria for the inclusion of data into the study were: (i) the reported K_p values plausibly represent the true steady-state distribution or the pseudo equilibrium and (ii) the V_{ss} and fu_p values in rats were available. It was assumed that all organs were non-eliminating such that the experimental and predicted K_p values were not affected by extraction ratio. The stereoselectivity was also considered; thus, the R and S enantiomers were regarded separately. In addition, experimentally determined LogP and pKa values were preferably used; if these were not available, calculated values were used [46,47]. As has been observed previously, the correlation between calculated and experimentally determined values is in good agreement [5]. When the tissue-to-plasma water (K_pu) parameter is reported, as in Rodgers et al., [8,9] the associated K_p was obtained by multiplying the values of fu_p and K_pu. If more than one experimental tissue K_p value was obtained for a single compound, the geometric mean was used.

Regression model development

A stepwise multiple linear regression analysis using R (i.e. language and environment for statistical computing) [48] was employed to develop a tissue-specific K_p prediction algorithm based on V_{ss}, LogP, the degree of ionization and fu_p. The important drug-specific parameters in the tissue distribution were incorporated to account for inter-tissue variation with the resulting structure:

Eqn. 2-8 $$LogKp_{tissue} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

Where β_0, β_1, β_2, β_3, β_4 are coefficients and x_1, x_2, x_3, x_4 are Log Vss, LogP, fup, fi, respectively.

Eqn. 2-8 is the largest model considered for each model. Smaller models were considered through stepwise regression. At each step of the stepwise regression analysis, a variable was either added or removed. The process was stopped when the fit yielded the greatest reduction in the Akaike information criterion (AIC) statistic [49]. The best regression equation for a tissue was determined such that it satisfied all of the selection criteria: (i) the equation resulted in the smallest AIC value in the analysis, (ii) the equation had the smallest sum of squared residuals, and (iii) the inclusion of a variable and the sign of its coefficient were reasonable (discussed below). In addition, to detect if the predictor variables were linearly related (i.e., the multicollinearity issue), the variance inflation factor (VIF) for each equation was screened. The VIF indicates the increase in the variance due to collinearity. A VIF value of 5 was used as the cut-off criterion [19,50]. If a multicollinearity problem was deemed to be present (i.e., VIF > 5), a given predictor was deleted and the next best equation was sought based on the AIC statistics.

The dataset was divided into two subsets. Subset A was comprised of moderate to strong bases ($pKa \geq 7.4$). Subset B consisted of acidic and neutral compounds, zwitterions, and weak bases

(pKa ≤ 7.4). For each tissue and each subset, the collected data was randomly divided such that 80% was used as the development set and 20% was used as the test set.

Evaluation of the obtained regression equations

The predicted K_p values (Pred) were plotted against the observed K_p values (Obs) for the test datasets of Subset A and Subset B. The adjusted coefficient of determination (Adjusted R^2) was used as a measure of the percentage of K_p variability that was explained by the predictor variables [50]. This measure represents the goodness of fit of each obtained equation. The precision of the obtained equation was assessed using the root mean square error (RMSE) (Eqn. 2-9), which ranks the precision of an equation:

Eqn. 2-9 $$RMSE = \sqrt{\frac{\sum_{1}^{n} (\log(\mathrm{Obs}_i) - \log(\mathrm{Pred}_i))^2}{n}}$$

Comparison of the accuracy of the model with the accuracy of the models developed by Jansson et al. and Rodgers et al.

Using the Subset A and Subset B test datasets, the accuracy of the algorithm was compared against the accuracy of an existing correlation-based [5] and a TCB model [8,9], both of which have been found to be good K_p predictors compared to other published algorithms [33,34]. The relative prediction accuracy was measured by calculating the percentage of predicted K_p values that exhibited a less than two-fold error deviation from the experimental data.

For each of the three algorithms, a measure of bias, the average fold error (AFE), was calculated (Eqn. 2-10). The AFE indicates an under-prediction (AFE < 1) or an over-prediction (AFE > 1) compared to the observed values. The absolute average fold error (AAFE) quantifies the overall

magnitude of the deviation between the predicted and the observed K_p values (Eqn. 2-11). To rank the overall precision of the model, the root mean squared error (RMSE) was calculated (Eqn. 2-9).

Eqn. 2-10 $$AFE = 10^{\left[\frac{1}{n}\sum_{1}^{n}\log\left(\frac{\mathrm{Pred_i}}{\mathrm{Obs_i}}\right)\right]}$$

Eqn. 2-11 $$AAFE = 10^{\left[\frac{1}{n}\sum_{1}^{n}\left|\log\left(\frac{\mathrm{Pred_i}}{\mathrm{Obs_i}}\right)\right|\right]}$$

2.4 Results

Development and prediction accuracy of the algorithm

The K_p prediction equations for moderate to strong bases (Table 2-2) and acids, neutrals and weak bases (Table 2-3) demonstrated a positive association between the V_{ss} and the observed K_p values (Figure 2-1). The V_{ss} parameter was used as a primary predictor of all tissue K_p values. The incorporation of LogP significantly improved the correlation between the tissue K_p values and the V_{ss} for the adipose and lung tissues. The fu_p was a key factor in the muscle K_p prediction (Tables 2-2 and 2-3). In the analysis of the heart, lung and muscle, the degree of ionization was an important predictor for all classes of compounds. No single equation displayed multicollinearity; thus, all of the VIF values were less than 5. For moderate to strong bases (Subset A), the degree of ionization had a positive effect on the K_p, whereas it had a negative effect on the K_p for Subset B.

Table 2-2. Correlations between the experimentally derived rat K_p values, the V_{ss} and the physicochemical parameters for strong to moderate bases (Training Set A)

Tissue	n	Regression parameters					Adjusted R^2	RMSE
		Intercept	LogV_{ss}	LogP	Fi	fup		
Adipose	33	-0.800	0.500	0.241	-	-	0.66	0.299
Bone	24	-2.157	0.86	-	2.122	-	0.68	0.263
Brain	47	-0.406	0.804	0.071	-	-	0.37	0.499
Gut	27	-5.191	0.711	-	5.672	0.275	0.68	0.236
Heart	50	-1.514	0.850	-	1.648	-	0.84	0.169
Kidney	54	0.405	0.861	-	-	0.309	0.53	0.308
Liver	52	0.392	1.035	-	-	-	0.48	0.415
Lung	51	-5.585	0.933	0.201	5.726		0.80	0.289
Muscle	53	-2.074	0.707	0.056	1.902	0.318	0.75	0.191
Spleen	9	0.066	1.041	-	-	-	0.84	0.159
Skin	28	-0.144	0.663	0.033	-	-	0.80	0.122

R^2, coefficient of determination; RMSE, root mean square error

Table 2-3. Correlations between the experimentally determined K_p values, the Vss and the physicochemical parameters for acids, weak bases and neutral compounds (Training Set B).

Tissue	n	Regression parameters					Adjusted R^2	RMSE
		Intercept	LogV_{ss}	LogP	Fi	fup		
Adipose	21	-0.298	1.144	0.231	-	-	0.64	0.374
Bone	13	-0.245	0.984	-		0.42	0.87	0.142
Brain	31	0.085	0.605	-	-0.832	-	0.67	0.302
Gut	26	0.043	0.831	0.067	-	-	0.62	0.238
Heart	35	0.146	0.644	-	-0.308	-	0.75	0.215
Kidney	31	0.463	0.425	-	-0.316	-	0.39	0.277
Liver	33	0.376	0.726	0.074	-0.333	-	0.79	0.237
Lung	32	-0.434	0.693	0.185	-0.286	0.520	0.83	0.222
Muscle	38	-0.122	0.65	-	-0.431	0.269	0.7	0.249

Spleen	18	0.136	1.008	-	-0.26	-	0.77	0.241
Skin	26	-0.331	0.544	0.158	-0.318	0.384	0.73	0.186

R^2, coefficient of determination; RMSE, root mean square error

Using the test datasets (Table 2-4), the calculated K_p values were in good agreement with the experimentally determined K_p values. Sixty-seven and sixty-two percent of the predicted K_p values fell within a two-fold deviation error of the experimental K_p values for Subset A and Subset B, respectively (Figure 2-2), which demonstrates similar relative prediction accuracy. Based on the RMSE values, the equations for moderate to strong bases had better precision (0.40) than those obtained for acids, neutral compounds and weak bases (0.44).

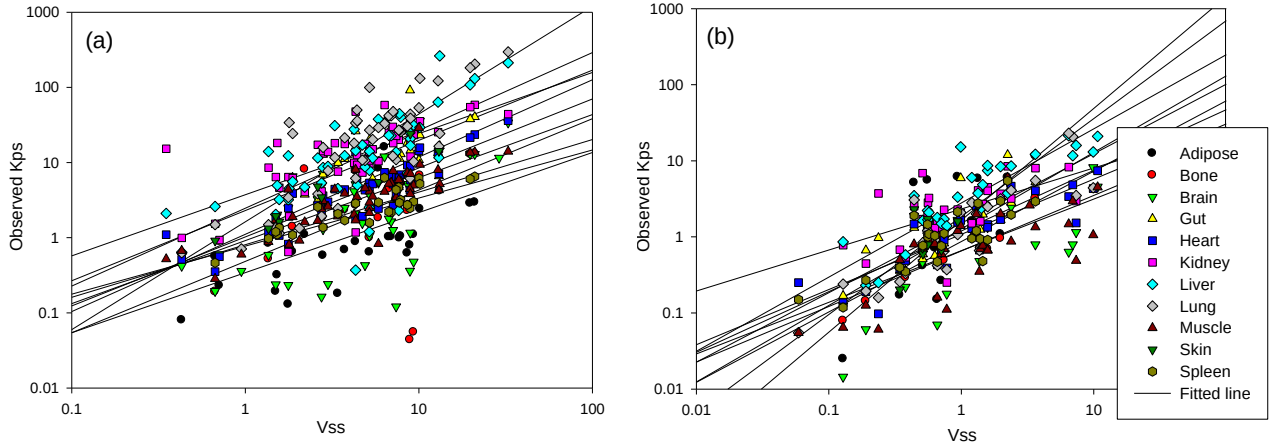


Figure 2-1. Association between V_{ss} and observed K_p values for (a) moderate to strong bases and for (b) acids, neutral compounds, and weak bases. The lines indicate the relationship between Vss and the observed K_ps for each tissue.

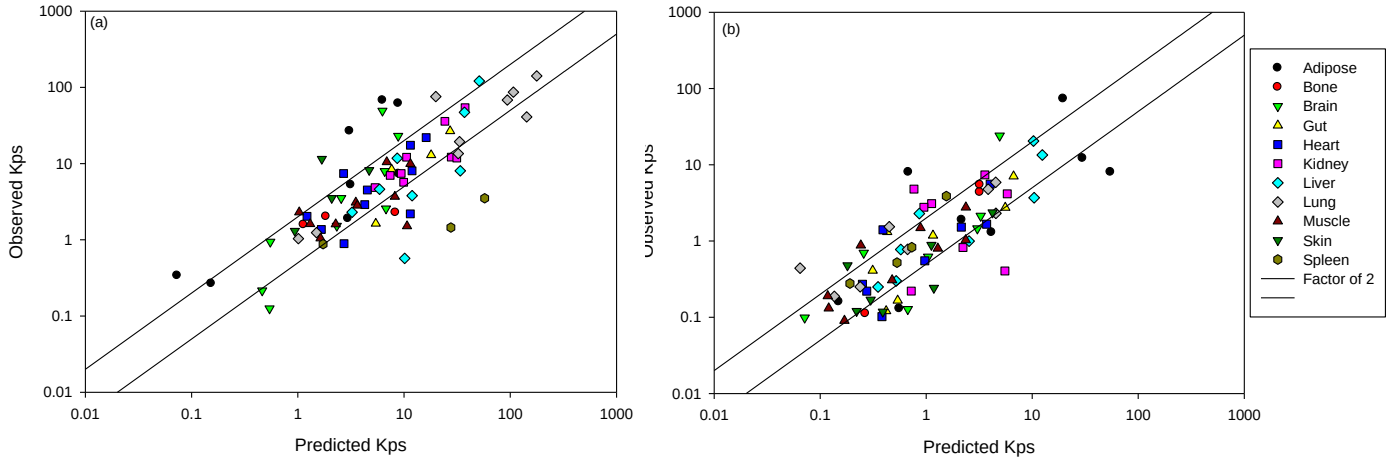


Figure 2-2. Logarithmic plot of observed vs. predicted K_p values for (a) moderate to strong bases (test set A) and for (b) acids, weak bases and neutral compounds (test set B). A total of 20 compounds and 154 tissue-specific K_p values are represented. The solid lines represent the ± 2-fold deviation from the experimental data.

Comparison of the K_p prediction accuracy of the proposed algorithm with the accuracy of published algorithms

The K_p values for the Subset A and B test datasets were predicted using the algorithms presented in this study as well as with the algorithms developed by Jansson et al. [5] and Rodgers et al. [8,9]. In terms of the overall prediction performance, the proposed model had greater predictive performance with lower RMSE values, AFE values closer to 1 and the greatest percentage of values within a 2- to 3-fold deviation error from the experimental values (Table 2-4). The prediction accuracy of the algorithms was tissue-dependent (Figure 2-3). For both Subsets, the presented algorithm had better prediction accuracy for the brain, kidney, liver, muscle and spleen K_p values. The adipose K_p values obtained with the proposed algorithm were under-predicted and had a poorer predictive accuracy compared to published algorithms. In addition, all algorithms resulted in a poor prediction of both the heart and the muscle K_p values for phencyclidine and FTY-720 in Subset A (see outliers in Figure 2-3).

Table 2-4. Accuracy of the K_p prediction obtained using the proposed algorithm and previously published models for the test datasets A and B [5,8,9]

	Model	n	AFE	AAFE	% within ± 2-fold of the experimental data	% within ±3- fold of the experimental data	RMSE
Test set A (Moderate to strong bases)	Proposed algorithm	77	1.04	1.99	67%	78%	0.40
	Jansson et al. [5]	72	0.67	3.12	53%	66%	0.67
	Rodgers et al. [8,9]	77	1.69	3.37	29%	51%	0.61
Test set B (Acids, neutral compounds, and weak bases)	Proposed algorithm	77	0.91	2.17	62%	75%	0.44
	Jansson et al. [5]	73	1.19	2.60	50%	67%	0.51
	Rodgers et al. [8,9]	77	1.49	3.40	53%	59%	0.72

AFE, average fold error; AAFE, Absolute average fold error; RMSE, root mean square error

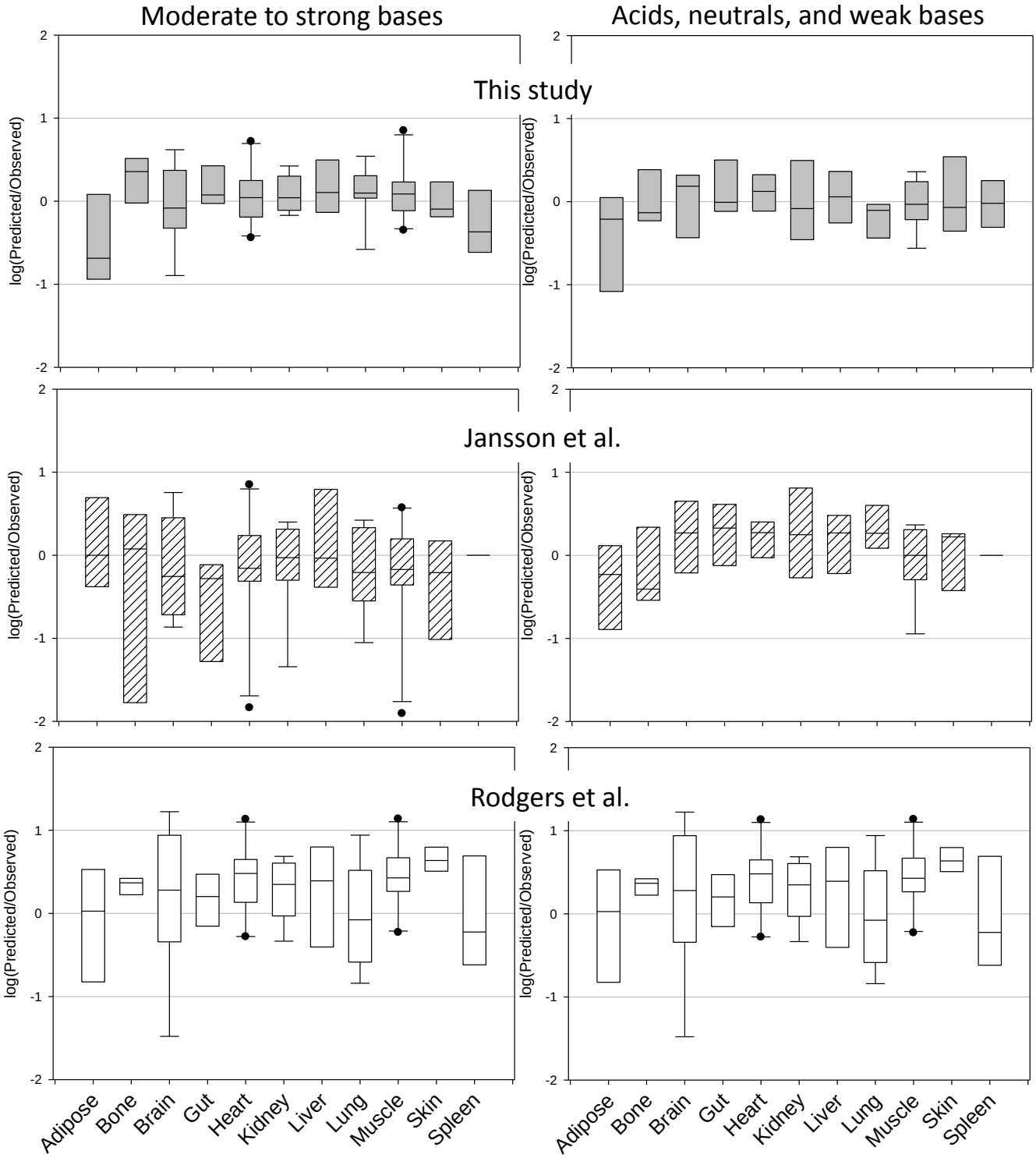


Figure 2-3. Box and Whisker plot of the logarithm of the ratio between the predicted and observed K_p values. The boxes represent the median (line) and the 25th and 75th percentiles; the bars represent the the 5th and 95th percentiles. The dots indicate the outliers.

2.5 Discussion

This study proposed a correlation-based K_p prediction algorithm that was built using a total of 96 compounds and 723 tissue K_p values. The relationships between the experimentally determined V_{ss} and the tissue K_p parameters, in addition to the physicochemical properties of the investigated drug, were used to derive the relevant K_p prediction equations. The algorithm differs from other correlation-based prediction algorithms due to its direct use of V_{ss} as a primary predictor variable and its use of the unbound fraction of the drug in the plasma and the degree of ionization as secondary predictor variables.

Our approach directly uses V_{ss} as a K_p predictor variable, whereas Jansson et al. [5] used the muscle K_p as a main predictor. In Jansson et al.[5], the muscle K_p can be derived from the V_{ss}; this derivation, however, can potentially cause great uncertainty in the estimated value of the muscle K_p. When the experimental muscle K_p value is used as an input in Jansson et al.'s model [5], a better prediction performance was observed (data were not shown in present study). However, the value of the muscle K_p is not likely to be available, which limits the use of Jansson et al.'s model [5]. By using the positive relationship between the tissue K_p values, an *in vivo* parameter (i.e., V_{ss}) and physicochemical descriptors (i.e., LogP, fup, and the degree of ionization), our method had better prediction accuracy than Jansson et al.'s model [5].

Moderate to strong bases often have large volumes of distribution with significant inter-organ variation [44]. One of the contributing factors to this variation is the uneven pH difference between the plasma and the tissues. Basic drugs tend to be stored in tissues with a pH that is lower than their pKa values. Due to the lower pH in the tissues, there would be a greater fraction of ionized species than unionized species and the positively charged ionized fraction would electrostatically interact with the negatively charged cell constituents. Even small differences in

the pH between the matrices and the plasma, which has a pH of 7.4, and tissues with a lower pH, such as the lung (pH 6.6), muscle (pH 6.81) and kidney (pH 7.22) (Table 2-1), are likely to create a large pH gradient that would result in the accumulation of a basic drug in a tissue [6]. The electrostatic interaction of the ionized fraction with acidic phospholipids, such as phosphatidylserine, phosphatidylinositol, phosphatidylglycerol and phosphatidic acid [8], is a crucial factor in the inter-organ variability of the tissue distribution [14]. There is a positive relationship between the K_p values and the concentration of acidic phosphatidylserine for moderate to strong bases that contain amines [44]. In addition, tissues vary in their acidic phospholipid composition. Thus, due to its inclusion as a predictor variable, the degree of ionization was expected to have a positive effect on the tissue partitioning for moderate to strong bases, which was indeed demonstrated in the resulting regression equations. The poor K_p prediction for some basic drugs can be explained by ion trapping. Basic drugs tend to be concentrated in lysosomes due to ion-trapping and/or intracellular binding. Unionized bases penetrate membranes and localize to acidic environments in cells, such as lysosomes. In an acidic organelle, bases become protonated and are thus unable to diffuse to the cytosol [51]. This behavior is an important factor in the drug distribution in lysosome-rich tissues, such as the liver, lung and kidneys [52]. Ion-trapping is the primary driving factor in the intracellular retention of hydrophilic strong bases, whereas ion-trapping and intracellular binding are equally important in the intracellular retention of polar strong basic drugs with high lipophilicity (e.g., propranolol) [52]. One of the outliers in our study was imipramine; there is a clear deviation between the observed and the calculated K_p values of this drug in the liver, lung, and kidney. Lysosomal trapping is responsible for approximately 10% of the distribution of this compound [53]. Because

the tissue pH that was used in the calculation of the degree of ionization was that of the whole tissue and not that of the individual organelles, this deviation is reasonable.

Neutral compounds, acids and weak bases (pKa ≤ 7) are likely to behave similarly to each other. In the plasma and tissues, these compounds primarily exist in their neutral form and only a small portion of these are ionized. In addition, hydrophobic interactions between the neutral components of a cell and reversible binding to extracellular proteins are expected to be prevalent with these compounds [8,9]. The level of tissue partitioning of weak bases is generally similar across the body and independent of the concentration of phosphatidylserine in the tissues [44]. The accumulation of acidic drugs, however, is a function of the differential pH between the plasma and the different tissues. The high degree of ionization of acidic drugs in the plasma would limit their entry into cells; in addition, once inside a cell, the acidic phosphatidylserine would have repulsive electrostatic interactions with the ionized fraction of these acidic drugs [6]. As a result, acidic drugs tend to accumulate to a greater extent in tissues with a higher pH because the unionized fraction in these tissues is greater than in tissues with a lower pH. Our study demonstrated that the K_p values of acidic drugs are negatively correlated with the degree of ionization (Table 2-3).

In general, the fu_p and V_{ss} parameters have a positive relationship. However, an increase in fu_p does not yield a proportional increase in V_{ss}, especially when a drug is found to be mostly bound to proteins [46]. This result indicates that the protein binding information is an important factor that should be utilized in the estimation of the tissue distribution of these drugs. Thus, fu_p provides information on distribution patterns that V_{ss} alone cannot convey. Despite the association of these variables, no mathematical evidence of collinearity was found in the construction of the prediction equations.

The prediction of the brain K_p is considered to be a challenge due to the blood brain barrier (BBB), which prevents many molecules from penetrating into the brain [8,54]. Tight junctions between the endothelial capillary and the glial processes near the capillaries make the BBB impermeable to polar molecules [55]. In general, lipophilicity has a positive effect on the drug partitioning to the brain because only lipophilic drugs can be transported through the BBB by simple diffusion. However, if the drug is a substrate of p-glycoprotein (Pgp), the resultant poor permeability of these lipophilic drugs may be the result of the efflux function of Pgp. Thus, the observed brain K_ps for Pgp substrates would account for additional processes, such as the rate of drug partitioning either by passive diffusion or by active transport, the rate of drugs that are repelled back to the blood by Pgp, and the non-specific binding to the BBB [56]. The presented approach assumes that the tissue partitioning is driven by the passive transport of a molecule into tissues, even though one group of researchers has questioned the validity of assuming passive diffusion for any drug [57]. The input K_p for a PBPK model is the K_p that assumes passive diffusion since active processes affecting permeability are accounted for separately. However, for algorithm development, the lack of consideration of active processes in the development datasets may have led to the poor prediction accuracy that was observed with the brain K_p (Figure 2-3). Although a poor brain K_p prediction with a relatively large standard deviation was obtained, the presented algorithm resulted in a better prediction of this parameter than other models.

An under-prediction was observed with the adipose K_p values for both test datasets. A poor prediction of the adipose tissue was also reported in the previous K_p estimation studies that used a correlation-based approach [4,5,7]. The possible reason for this decreased accuracy in the adipose K_p prediction is the different lipid composition of this tissue compared to other tissues. In

adipose tissue, neutral lipids are more abundant than other cell constituents, such as phosphatidylserine, and other lipids [7,8,29]. Therefore, in these cases, the hydrophobic interactions are more dominant than the electrostatic interactions, thereby leading to the accumulation of lipophilic drugs in adipose tissues. Jansson et al. [5] stated that the adipose K_p, prediction from the muscle K_p was less accurate compared to other tissues. Poulin and Theil presented a different approach that used an adjusted skin K_pu to estimate the adipose K_pu [7]. It has been suggested that the variation in the adipose tissue K_p among the different classes of drugs cannot be simply explained by the physicochemical and *in vivo* parameters. Therefore, a different approach is required to increase the prediction accuracy of the adipose K_p. Another contributing factor in the poor prediction of the adipose K_p may be the inaccuracy of the LogP values and to the inter-laboratory variation that exists in the determination of these parameters [5]. Thus, this result highlights the importance of using accurate physicochemical information for the prediction of K_p.

Phencyclidine is a cationic-amphiphilic drug that acts mainly on the inotropic glutamate receptors in the rat brain [58]. All three algorithms resulted in a poor K_p prediction for this drug, especially in the muscle and heart (shown in Figure 2-3 as an outlier). The other outlier for heart K_p was FTY-720, which is a therapeutic drug used for the treatment of heart failure through the activation of Pak1 signaling [59]. Both of these drugs are highly lipophilic with LogP values greater than 4.00 and are highly ionized at physiological pH. Because the values of some inputs, such as LogP and V_{ss}, were large, the algorithms yielded larger K_p values compared to the experimentally determined K_p. There is no current explanation for these results since other tissues within each compound were adequately described. A possible explanation is the presence of an efflux transporter in those affected tissues that was not considered.

Correlation-based models, unlike TCB models, are dependent on the dataset that is used in their derivation. The mechanistic equations are potentially applicable for any species if the tissue-specific physiological parameters are available. Parameterizing TCB models requires less *in vivo/ex vivo* (e.g., fu_p) information than correlation-based models, which require muscle K_p, V_{ss} or RBC_u. TCB models require complex parameterization. Many researchers have strived to develop prediction algorithms using complex parameters to describe the distribution process at a cellular level within a mechanistic structure. Some K_p prediction algorithms require many input parameters, such as the blood-to-plasma ratio, red blood cell partitioning data, and the phosphatidylcholine-to-water partition coefficient at pH 7.4 [6-8], that may be unavailable. Furthermore, some TCB algorithms are mathematically heavy and their reproduction is difficult. However, correlation-based models rely on the dataset [8]. If the dataset used is small, the data pool may not represent an accurate sampling and the fit is thus likely to be sensitive to the inclusion/exclusion of an observation. The input parameters (e.g., muscle K_p [4,5,7], skin K_p [7], adipose K_p [4], and RBCu [7]) are often not easily obtained, which limits the ability to make a priori predictions. The proposed algorithm was derived using a larger dataset than all previously developed correlation-based algorithms.

2.6 Conclusion

The derived K_p prediction algorithm is mathematically simple and employs input parameters generally available in pre-clinical drug development or early toxicological assessment. In addition, the model has greater prediction accuracy in comparison to the best correlation-based and TCB models that are currently available.

Chapter 3

Development of a decision tree to classify the most accurate tissue to plasma partition coefficient algorithm for a given compound in rats

3.1 Introduction

Partitioning of a compound into a tissue is a complex process. In PBPK modeling, the estimation of a compound's distribution parameters has limited the accessibility of this modeling technique due to difficulties in their experimental determination (i.e. K_ps) in the species of interest [28]. In order to overcome this barrier, numerous *in silico* methods for K_p prediction have been developed [2,4-9,12,13,16]. Despite increasing attention and interest in the accurate prediction of compound distribution data or tissue dosimetry profiles, a standard K_p prediction method has not yet been determined. There is no single prediction algorithm that is applicable for all compounds in all tissues (see Table 3-1). The accuracy of the pre-existing K_p prediction algorithms still require improvement [6]. The predictability of any single K_p prediction algorithm, whether it is a tissue composition based or a correlation based algorithm may vary depending on the physico-chemical properties of a compound and/or the physiological parameters of an organism. These algorithms may also have varying tissue specific prediction accuracies. Furthermore, the experimental determination of all of the required compound specific chemical descriptors and *in vitro* and *in vivo* input parameters can limit the use of some K_p prediction algorithms. In other words, the availability of these parameters often determines the usability of an algorithm. For these parameters, estimation equations are suggested as an alternative to experimental

determination. Therefore, the estimation equations will allow use of K_p prediction algorithms with a minimal number of readily available compound specific parameters. With the use of estimation equations, this study aims to determine the best performing algorithm in a specific physico-chemical space for a single tissue. In order to address this problem, statistical classification techniques are used.

Machine learning methods for decision tree development

Machine learning refers to the construction of a system that can learn from training data. Learning algorithms for classification learn based on certain data (e.g. measurement data or categorical data) and a response of interest [60]. The objective of machine learning is to characterize the observed phenomenon and generalize it (i.e. inductive inference), in an attempt to make accurate predictions for a new sample [60]. Decision tree learning is a decision support system that uses a tree-like model of decisions. The decision tree based classification methods were investigated to identify the best performing algorithm in a specific physico-chemical space for each tissue.

Table 3-1. Summary of applicability of K_p prediction algorithms

	Algorithms	Acid	Base	Neutral	Zwitterion	Adipose	Bone	Brain	Gut	Heart	Kidney	Liver	Lung	Muscle	Pancreas	Skin	Spleen	Testes	Thymus	RBC
1	Bjorkman [4]	v	v			v	v	v	v	v	v	v	v			v				
2	Berezhkovskiy [12]	v	v	v	v	v	v	v	v	v		v	v	v		v				
3	Rodgers et al [8]		v			v	v	v	v	v	v	v	v	v	v	v	v		v	
	Rodgers & Rowland [9]	v		v	v	v	v	v	v	v	v	v	v	v	v	v	v		v	v
4	Schmitt [6]	v	v	v	v	v	v	v	v	v	v	v	v	v	v	v	v	v		v
5	Jansson et al [5]	v	v	v	v	v	v	v	v	v	v	v	v	v		v				
6	Poulin &Theil [7]		v			v	v	v	v	v	v	v	v	v		v	v		v	
7	Yun and Edginton [19]	v	v	v	v	v	v	v	v	v	v	v	v	v		v	v			
8	The proposed study	v	v	v	v	v	v	v	v	v	v	v	v	v		v	v			

Recursive partitioning method

The recursive partitioning method creates a decision tree that aims to correctly categorize members of groups based on several variables [61]. The variables in this analysis are not assumed to follow any specific statistical distribution. A classification tree is represented as an inverted tree with a root node at the top, branches connecting nodes and leaves at the bottom [50]. The schematic below presents an example of an output of the recursive partitioning method. At each node, a question regarding a variable is posed. The leaves denote classifications (i.e. a K_p prediction algorithm) and the child nodes represent splits that lead to the classifications. The numbers at the end of a leaf (Figure 3-1) depict the number of cases within a test dataset that were best represented by different categories or K_p prediction algorithms. For the leaf in Figure 3-1, the classification is category 2 because it has the highest frequency in the leaf (Figure 3-1).

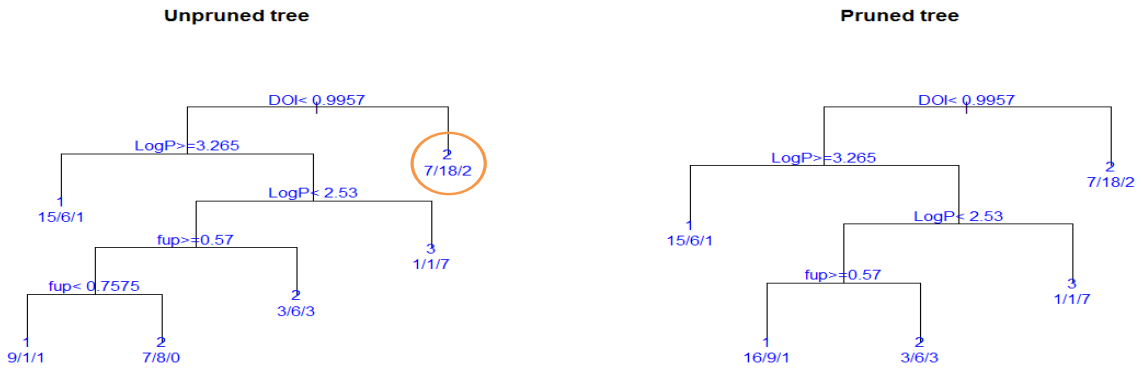


Figure 3-1. An example of a classification tree developed using recursive partitioning. The left tree is unpruned whereas the right tree is pruned.

The classification tree is built using the following steps. A variable (e.g. LogP) that best splits the data into two groups is based on the criterion of the Gini index (Eqn. 3-1). Let I(A) be an impurity function of a node A.

Eqn. 3-1 I(A)= $\sum_{k \neq j} p_k \cdot p_j = 1 - \sum_j p^2{}_j$

where p_k is the fraction of samples in a node A that belong to class k (k=1,2,…K). The probabilities (i.e. p_k, p_j) are calculated from node frequency (e.g. Figure 3-1 - 7/18/2). A split is chosen when the split results in maximal impurity reduction. At each possible split, the sample is divided into child nodes [62]. The data is subdivided repeatedly until there is no reduction in impurity of a node is possible. If a case with the response "true" to the question posed, it is sent to the left child node and the "no" responses are sent to the right child node.

The schematic (Figure 3-1) shows an example of a pruning process of the recursive partitioning method. Large trees use a larger number of variables, and these trees may result in overfitting of the data. In order to avoid this, a cost-complexity pruning is performed to extract insignificant splits [63]. The aim of the tree pruning is to identify a nested version (i.e. subtree) of a fully grown tree so that the nested tree minimizes the measure of cost-complexity on an independent test set [63]. The cost-complexity measure can be expressed as following:

Eqn. 3-2 $R_\alpha(T) = R(t) + \alpha|T|$

where $R_\alpha(T)$ is the misclassification cost of the whole tree at a complexity parameter α, and R(t) is the misclassification cost evaluated at the node. The number of nodes is denoted as $|T|$. A complexity parameter α (α > 0), which penalizes cost, is assigned a one unit increase in complexity (i.e. addition of a terminal node) [64].

The sum of all misclassification costs is converted into a penalty for the complexity of the tree. The complexity of the tree increases as the number of nodes (i.e. size of the tree) increase because the data is further divided into the smaller parts. The complexity parameter α adjusts for the influence of tree size on cost-complexity. If α = 0, the largest tree will be chosen. If α approaches infinity, then a root node without any child node will be selected ($|T| = 1$).

Due to the absence of an independent test set in most cases, cross validation is used as an alternative to external validation. Recursive partitioning is implemented in the *rpart* package in R and by default, the *rpart* function in *rpart* package performs 10 fold cross validation [48,64,65]. In this procedure, the dataset is divided into 10 equally sized segments. Nine segments are used for growing a classification tree and the tenth segment is used as a test set. To obtain the optimal tree, the complexity parameter that minimizes the 10 fold cross validation error is selected.

The function *prune()* in the *rpart* package [48,65] trims the tree to the complexity parameter value that minimizes cross validation error [64]. For the left tree in Figure 3-1, according to complexity parameter, it is found that a tree with 4 splits had a lower cross validation error as compared to a tree with 5 splits. As a result, the last split was extracted.

Random forest and bootstrap aggregation

Random forest and bootstrap aggregation (Bagging) are also methods of classification. These methods are based on a collection of classification trees instead of a single tree, as in recursive partitioning. These methods generate multiple versions of a classification tree by using bootstrapping, and aggregate the classification from the various trees. Bootstrapping [66] is a procedure inherent in both random forest and Bagging. This procedure determines the reliability of estimates in a statistical analysis by generating resamples of the original dataset with the same

sample size [50]. If the dataset set follows the assumption of independent and identically distributed observations, a bootstrap sample is drawn with the same sample size as the original dataset with replacement.

- *Random forest*

A random forest is defined as a classifier that is comprised of a set of classification trees

Eqn. 3-3 $\{h(x,\Theta_k), k = 1,.....N\}$

where **x** is an input vector (i.e. explanatory variables) and the (Θ_k) are the independent identically distributed random vectors [67]. *N* bootstrap samples are drawn from the training data. For each bootstrap sample, the number of input parameters, m_{try} (m_{try}=1,2,…M), are randomly chosen (m_{try} << M) and a classification tree is grown in the same way as recursive partitioning. In other words, each tree is created using a random set of samples and input parameters.

At each node of a tree, the variable that results in the greatest decrease in impurity is selected to separate the child nodes. Much like in recursive partitioning, the impurity of the node is measured by the Gini index (Eqn. 3-1). The splitting continues until the child node has only samples that belong to the same class.

Each tree is grown without pruning, in that the tree is grown to its largest extent and the tree size is not optimized. The random selection of variables results in trees with minimal correlation to each other. In order to classify an object from input x (Eqn. 3-3), the object (x_{new}) is put to each of the trees grown in the forest, and consequently, each tree classifies it to a group. With a new input x_{new}, each tree results in a classification. Among unpruned trees (e.g. by default n_{tree} = 500), the classification with the most votes is selected by the forest. For each tree, about 67% of the data is drawn from an original dataset to create a tree by recursive partitioning, as described above. The remaining 33% of the data is left out as an 'out-of-bag' (OOB) sample [68]. As

bootstrap samples are drawn with replacement, about 36 % of the total data is OOB on average [69].

Each classification tree created from a training set makes predictions for the OOB sample at each iteration. From the aggregated OOB predictions, the OOB estimate of error rate is calculated. [69] This internal estimate of error rate tends to overestimate the error that a tree grown from the total dataset would. However, it does allow for an assessment of the classification performance of a random forest. Via a built-in cross validation function of *rfcv()* in the *randomForest* package in R [48,69], the random forest can be tuned by using the optimal value of the number of variables (m_{try}) [70].

- *Bootstrap aggregation*

In the Bagging method [71], a set of classification trees are grown. A training data α is comprised of {(yi, xi), i=1…I} where y is class and x is input vector. A classification tree from the training dataset can be expressed as φ(x, α). *N* bootstrap samples α_k (k=1….N), is drawn from a training set α at random, but with replacement. A decision tree without pruning is grown based on each bootstrap sample using recursive partitioning. However, unlike the random forest method described above, all variables are considered as a potential split for each tree (m_{try}=M) [71]. Due to random variation inherent in bootstrapping, each tree differs from one another. A set of classification trees φ(x, α_k) is aggregated by the majority vote as the same principle of majority vote in the random forest method [69,72]. Both random forest and bagging methods exploit the fact that a single classification tree is very unstable and that a small change in the training set can result in different classification. But, the aggregation of multiple versions of the classification trees yields a better prediction [61].

3.2 Objectives and Hypothesis

The current study aims to develop a decision tree that will choose the most accurate algorithm for the prediction of tissue specific K_ps. This study employed a classifier learning algorithm to develop a classification tree that will identify the most precise algorithm for a compound within a given physico-chemical space. The objectives of the predictive classifier are: (i) to provide K_p predictions using readily available parameters and (ii) to use the most accurate prediction algorithm to calculate tissue-specific K_ps for a compound. It is hypothesized that the developed classification tree(s) will produce a more accurate overall prediction of K_ps than any one K_p prediction algorithm alone.

3.3 Methodology

Data collection

A database of experimentally derived partition coefficients with corresponding compound physico-chemical properties were created from the literature using several MEDLINE searches. *In vivo* parameters such as the fraction unbound in plasma (fup) and volume of distribution (Vss) were also included in the database. Data was included in the study based on the following criteria: (i) reported K_p values plausibly represent true steady state distribution/ pseudo equilibrium and (ii) fup, pKa, and one of the lipophilicity measures (i.e. LogP, LogD, LogKvo:w) were available. When experimental physicochemical parameters (e.g. all lipophilicity measures, pKa) were not available in the literature, the values were obtained from predictions made in ChemEbl [47]. Experimentally determined values were preferably used over predicted values. Stereoselectivity of a compound was considered, if applicable, so that R and S enantiomers were considered separately. As shown in Table 3-1, decision trees for pancreas,

testes, thymus and RBC were not generated since the number of data points was insufficient for a classification analysis.

Estimation of required inputs

Table 3-2 presents the required input parameters for each algorithm. In the event that a required input parameter was not available, it was calculated based on regression equations presented in Table 3-3. For example, if only LogP was available but LogD was the necessary input parameter, LogD was calculated using equations based on the equations derived by Poulin et al [15] (see Table 3-3). For some input parameters [e.g. LogMA, LogHSA, and blood: plasma ratio (B:P)], a regression equation was derived using the datasets in the Rodger *et al.* [8] and Schmitt [6] publications.

Affinity for blood cells (K_puBC) (i.e. unbound compound concentration in blood cells) is one of the required parameters for the Rodgers *et al.* [8,9] algorithms. K_puBC is the function of fup, B:P and hematocrit. K_puBC is estimated using the standard equation (Eqn. 3-10) in the Rodgers models [8,73]. In the absence of an observed B:P, B:P is estimated using the estimation equation (Eqn. 3-11) proposed by Paixao *et al.* 2009 [74]. This equation was derived from Rodgers *et al.* 2006. The assumptions for the equations are that: (i) in erythrocytes, there is no extracellular space and (ii) albumin and lipoproteins are not contained within the space.

While the first approach to B:P estimation was the use of a mechanistic model as descried above, another approach was also taken for B:P estimation. This was the development of a regression equation (Eqn. 3-12). Experimentally determined B:P, LogP and fup (n = 28) were obtained from Rodgers *et al.* [8] and a predictive regression equation was developed based on the dataset. For the linear regression analysis, the statistical software R version 2.12 [48] was used. The estimation

equation that yielded a more accurate K_pu prediction when compared to the observed K_pu values was selected for the calculation of K_ps for Rodgers *et al*. [8,9] in this study.

For the calculation according to Schmitt's algorithm [6], the logarithmic value of phosphatidylcholine: water partition coefficient at pH 7.4 (LogMA) and the logarithmic value of human serum albumin (LogHSA) must be estimated in the absence of the experimentally determined values. Using the dataset provided by Schmitt, LogP, LogMA, and LogHSA (n =60 data points) were obtained. The regression equations for LogMA (Eqn. 3-8) and LogHSA (Eqn. 3-9) were generated.

Table 3-2. Summary of K_p prediction algorithm and their main inputs.

Algorithm	**Approach**	**Main inputs**
Bjorkman [4]	Correlation based	Muscle K_p
Berezhkovskiy [12,16,29]	Tissue composition based	LogP, LogKvo:w, fup
Rodgers *et al.* [8,9]	Tissue composition based	LogP, pKa, fup, B:P
Schmitt [6]	Tissue composition based	LogP, LogD, LogKvo:w, LogMA, LogHSA, pKa, fup
Jansson *et al.* [5]	Correlation based	Vss, Muscle K_p, LogP, LogD, LogKvo:w
Poulin and Theil [7]	Correlation based	Muscle K_p or RBCu
Yun and Edginton [19]	Correlation based	Vss, LogP, pKa, fup

Table 3-3. Summary of equations used to estimate an unknown input parameter.

	Parameter	Description	Equation	Reference
Eqn. 3-4	Fut_lean tissue	Fraction of unbound compound in lean tissue	1/(1+(((1-fup)/fup)*0.5))	[2]
Eqn. 3-5	Fut_adipose tissue	Fraction of unbound compound in adipose tissue	1/(1+(((1-fup)/fup)*0.15))	[2]
Eqn. 3-6	LogD	Partition coefficient of octanol and water at specific pH	Monoprotic base $LogP - Log(1+10^{pKa1-7.4})$ Diprotic base $LogP - Log(1+10^{pKa1-7.4}+10^{pKa1+pKa2-2\times7.4})$ Monoprotic acid $LogP - Log(1+10^{7.4-pKa1})$ Diprotic acid $LogP - Log(1+10^{7.4-pKa1}+10^{2\times7.4-pKa1-pKa2})$ Zwitterions $LogP - Log(1+10^{pKa\_base-7.4}+10^{7.4-pKa\_acid})$ Where pKa1 > pKa2	[15,16]

Eqn. 3-7	LogKvo:w	Logarithmic value of partition coefficient between vegetable oil and water.	1.115*LogP-1.34	[75]
Eqn. 3-8	LogMA	Logarithmic value of membrane affinity.	LogMA =1.294+0.304*LogP This equation was obtained using Schmitt's dataset. In the dataset, there were 60 logMA values available. The regression equation was developed and was statistically significant (P<0.05).	[6]
Eqn. 3-9	LogHSA	Logarithmic value of Human serum albumin(HSA)	LogHSA=0.294+0.135*LogP This equation was obtained using Schmitt's dataset. In the dataset, there were 60 logHSA values available. The regression equation was developed and was statistically significant (P<0.05).	[6]
Eqn. 3-10	K_pu_BC (Affinity for blood cell)	Red blood cell to plasma partition coefficient of unbound compound, Affinity of a compound for a red blood cell	$\frac{BP-(1-Hematocrit)}{Hematocrit^{*}\ fup}$	[73]
Eqn. 3-11	K_puBC		$\frac{X\cdot f_{IW\_RBC}}{Y}+\left(\frac{Pf_{NL,RBC}+(0.3P+0.7)f_{NP,RBC}}{Y}\right)$ Where	[74]

			f_{IW}=.0914,f_{NL}=0.0017,f_{NP}=0.0029 For monoprotic base: $X=1+10^{pKa-7.22}$, $Y=1+10^{pKa-7.4}$ For monoprotic acids: $X=1+10^{7.22-pKa}$, $Y=1+10^{7.4-pKa}$	
Eqn. 3-12	Blood to plasma ratio(B:P)		Log(B:P) = -0.004282 +0.067028 LogP + 0.214590 Log(fup) (n=28 , R^2=0.40) This equation was obtained using Rodgers *et al.* [8] dataset. In the dataset, there were 28 experimentally determined BP values available. The regression equation was developed and was statistically significant (P<0.05).	[8]
Eqn. 3-13	Muscle K_p		$Vss = V_{plasma} + \sum_{1}^{n} V_{tissue,i} \times 10^{a\times\log(Kp,muscle)+b\times\log(lipophiliciy)+c}$	[5]
Eqn. 3-14	Degree of ionization at a tissue pH		$fi = 1-[1+10^{pKa-pH\ tissue}]^{-1}$ for monoprotic bases $fi = 1-[1+10^{pKa-pH\ tissue}+10^{pKa_1+pKa_2-pH\ tissue\times 2}]^{-1}$ for diprotic bases $fi = 1-[1+10^{pH\ tissue-pKa}]^{-1}$ for monoprotic acids $fi = 1-[1+10^{pH\ tissue-pKa}+10^{pH\ tissue\times 2-pKa_1-pKa_2}]^{-1}$ for diprotic acids $fi = 1-[1+10^{pka_{base}-pH\ tissue}+10^{pH\ tissue-pKa_{acid}}]^{-1}$ for zwitterions	[19]

Separation of classifier groups

For researchers requiring K_p prediction for a novel compound, the availability of input parameters will not be consistent. For example, when *in vivo* work has not been done on the compound, researchers are likely to have only physico-chemical input parameters and lack any *in vivo* input parameters such as muscle K_p. Therefore, a decision tree incorporating algorithms that require *in vivo* inputs will not be useful for the researcher. Based on this, several versions of the classification trees were created and were based on the likely groupings of input parameters researchers may have. Any additional algorithm-specific input parameters that were required were estimated using the equations in Table 3-2.

The development and evaluation of Classification tree #1 was dependent on compounds for which muscle K_p, one of the lipophilicity measures (e.g. LogP), pKa, and fup were available (Table 3-4). The development and evaluation of Classification tree #2 was dependent on compounds for which Vss, one of the lipophilicity measures, pKa and fup were available. The development and evaluation of Classification tree #3 was dependent on compounds for which one of the lipophilicity measures, pKa and fup were available. The algorithms that were classified in each of the Classification trees are listed in Table 3-4 along with the number of compounds used in the development and evaluation of each tree.

Table 3-4. Physicochemical and/or *in vivo* parameter inputs for a classifier algorithm and included algorithms for each group.

	Inputs for classification	Algorithms
Group 1	Muscle K_p, LogP,	Berezchkovskiy [12]
(N=107 compounds)	fi,	Bjorkman [4]
	fup, Class[a]	Rodgers *et al.* [8,9]
		Schmitt [6]
		Jansson et al. [5]
		Poulin and Theil [7]
Group 2	Vss, LogP,	Berezchkovskiy [12]
(N=97 compounds)	fi,	Rodgers *et al.*[8,9]
	fup, Class [a]	Schmitt [6]
		Jansson et al. [5]
		Yun and Edginton [19]
Group 3	LogP, fi,	Berezchkovskiy [12]
(N=121 compounds)	fup, Class [a]	Rodgers et al. [8,9]
		Schmitt [6]

[a]Class: acid-base properties of a compound (A: acid, B: base (pKa≥7.4), WB: base (pKa≥7.4), Z: zwitterion)

K_p calculations according to the previously published algorithms

To ensure that the use of estimated input parameters as defined in Table 3-3 produced K_p predictions that were similar to those predicted using existing algorithms, a comparison of outcomes was completed. K_ps were calculated according to each published equation using only those input parameters required for Classification trees #1 through #3 and using estimation equations for any remaining inputs required. For Rodgers *et al.*'s method, K_ps of bases with $pKa \geq 7$ were calculated by Rodger *et al.* [8]. LogKvo:w and B:P were estimated by Eqn. 3-7, Eqn. 3-12 (Table 3-2). K_ps of acids, neutrals, and weak bases were calculated by Rodgers *et al.* [9]. In Jansson's algorithm [5], K_p prediction equations of bases and neutrals, and K_p prediction equations of acid and zwitterions were separately used. For Classification Tree #1, the experimentally derived muscle K_p value was used as an input. For Classification Tree #2, experimental Vss was used as a direct input for those algorithms requiring it and was used to estimate muscle K_p in those algorithms where muscle K_p was an input (Eqn. 3-13). LogD and LogKvo:w were calculated as a function of LogP using Eqn. 3-6 and Eqn. 3-7. In Schmitt's model [6], compound class was separated by acids, neutrals, bases, and zwitterions and K_ps were calculated accordingly. LogMA and LogHSA were estimated using the regression equations Eqn. 3-8 and Eqn. 3-9. In the Yun and Edginton algorithm [19], K_ps were estimated by using equations for moderate to strong bases and equations for acids, neutrals and zwitterions. The degree of ionization at a specific tissue pH was calculated using Eqn. 3-14. Since Poulin and Theil's K_p prediction approach [7] was targeted for predicting K_ps for bases, only K_ps of bases were estimated. In Bjorkman's model [4], K_p prediction equations for acids and bases were separately developed and K_ps were calculated accordingly.

The difference between calculated K_p values using both experimental and estimated input parameters were compared to the calculated K_ps published in Rodgers *et al*. [8,9], Schmitt [6], and Jansson *et al*.[5]. The comparison could not be made for Berezhkovskiy [12], Bjorkman [4], and Poulin and Theil [7] as the calculated K_ps were not presented in their publications.

Mean fold error (MFE, Eqn. 3-16), average fold error (AFE, Eqn. 3-18), absolute average fold error (AAFE, Eqn. 3-19), and root mean square error (RMSE, Eqn. 3-20) were used to measure the deviance of the published algorithm predicted K_ps and the K_ps calculated using experimental and estimated inputs (Table 3-5).

Table 3-5. Statistics for comparative assessment of prediction accuracy

	Metrics	Formula
Eqn. 3-15	Fold Error (FE)	$\frac{\text{Pred}_i}{\text{Obs}_i}$ Where Pred_i is predicted value, Obs_i is observed value.
Eqn. 3-16	MFE	$\sum_1^n \left(\frac{\text{Pred}_i}{\text{Obs}_i}\right)$
Eqn. 3-17	% within k-fold error	$\left[\frac{1}{n}\sum_{i=1}^{n} I\left(\frac{1}{k} \le \frac{\text{Pred}_i}{\text{Obs}_i} \le k\right)\right]\times 100\%$, I(·) is an indicator function, k= 1.25,1.5,2,3
Eqn. 3-18	AFE	$10^{\left[\frac{1}{n}\sum_1^n \log\left(\frac{\text{Pred}_i}{\text{Obs}_i}\right)\right]}$
Eqn. 3-19	AAFE	$10^{\left[\frac{1}{n}\sum_1^n \left\lvert \log\left(\frac{\text{Pred}_i}{\text{Obs}_i}\right)\right\rvert\right]}$
Eqn. 3-20	RMSE	$\sqrt{\frac{\sum_1^n (\log(\text{Obs}_i) - \log(\text{Pred}_i))^2}{n}}$

Dataset Development

Using the compound specific properties and the *in vivo* parameter data in Group 1, 2, and 3 (Table 3-4), a comparison of experimentally derived K_ps with predicted K_ps from each applicable algorithm were made. The K_p prediction algorithm that resulted in a value that was closest to the experimental one was selected for the compound. The selected model for the compound was then coded numerically so that the compound could be categorized by the best predicting model (coded as in Table 3-6). This coded information was used as the dependent variable in the statistical analysis. In order to determine which K_p prediction method should be used for a given physicochemical space, statistical methodologies such as 'recursive partitioning method', random forest, and bagging were investigated in this study. A classification learning algorithm that identified the best prediction K_p algorithm with a lower classification error rate was chosen for this study.

Recursive partitioning and Classification learning algorithms

The recursive partitioning, bagging, and random forest methods were utilized to build a classifier that identified the most accurate K_p prediction model. Those classification analyses were performed using the statistical software R (version 2.14) [48]. Recursive partitioning is implemented in the *rpart* package. After an unpruned classification tree was grown, by using the function of *printcp()*, the cross-validated prediction error for different numbers of splits was calculated. A tree was pruned by setting the complexity parameter that resulted in the smallest cross-validation error.

Random Forest is implemented in *randomForest* package (4.6-6) [48,65]. Initially, the parameters were set to the number of trees in a forest (ntree= 500) and number of variable ($m_{try} = \sqrt{M}$) by

default. By using rfcv function embedded in the *randomForest* package [48,69], the optimal m_{try} that resulted in the smallest cross-validated error was chosen. A final random forest model was generated by setting the optimized variable of m_{try} when trees are grown. The Bagging function is implemented in the *ipred* package [48,72]. In this analysis, unpruned classification trees were grown from 25 bootstrap samples. The prediction of a new observation is aggregated by the majority vote [72].

Evaluation of classification performance of random forest, bagging and recursive partitioning

In order to find the most appropriate classification method, the output of 3 methods: random forest, bagging and recursive partitioning were compared. Using the same development dataset of n=99 (80% of the total dataset), tissue specific classification trees using recursive partitioning, bagging and random forest were generated. The sample R-code is shown in the Appendix 5. The rate of correct classification was used as a metric to determine which classification method performed best within this study. The rate of correct classification of each method was obtained using an independent test set of n = 23 compounds (20% of the total dataset). The classification method that resulted in the highest rate of correct classification in the most tissues was chosen for this study (Eqn. 3-21).

Eqn. 3-21 Rate of correct classification = $\frac{1}{n}\sum_{1}^{n} I(Obs_i = \mathrm{Pred_i})$

Where I(•) is an indicator function, Obs_i is observed classification, $Pred_i$ is predicted classification, and n is the number of observation.

Evaluation of the random forest using cross validation

The developed random forests for Classification tree #1, Classification tree #2 and Classification tree #3 that corresponded to each group in Table 3-4, were evaluated. The predictive performance of each Classification tree was evaluated with the total dataset by using 20 fold cross validation [70]. This method assumes that a random forest developed from 95% (19/20) of a total dataset is reasonably the same as a final random forest that is developed using 100% of the total dataset. The sample R-code is shown in the Appendix 6.

The steps taken in the 20 fold validation and analysis were as follows:

(i) The total dataset was partitioned into 20 subsets.

(ii) A random forest was created using a training set comprised of 19 subsets. The developed random forest then predicted the classification for samples in the 20^{th} subset as a test set. The predicted classification (e.g. best algorithm for compound X = Jansson *et al.*[5]) for the test set was recorded. This step was repeated 20 times so that each subset was used only once as a test set. As a result, each compound was used once as a test compound.

(iii) For the test dataset that includes all compounds, each compound is associated with a random forest generated best prediction algorithm.

(iv) The rate of correct classification is calculated (Eqn. 3-21).

(v) The K_p is calculated using the algorithm identified as the most accurate during the cross-validation (Table 3-6).

Using this method, the predictive performance of previously published algorithms was compared to the random forest generated K_ps with the use of the same total dataset (n=122 compounds, shown in Appendix 7, Appendix 8)

Table 3-6. An example of a dataset for the random forest analysis and corresponding calculated K_p values.

Compound	Observed Heart K_p	1.Berezhkovskiy [12]	2.Rodgers *et al.* [8,9]	3.Schmitt [6]	4.Jansson *et al.* [5]	5.Yun and Edginton [19]	Code1[a]	Code2[b]	Predicted K_p by a random forest
Compound1	3.87	5.74	8.18	27.59	14.84	**4.39**	**5**	5	4.39
Compound2	5.71	1.41	7.24	22.07	**5.22**	8.99	**4**	4	5.22
Compound3	2.61	1.02	0.72	1.64	**2.74**	3.62	**4**	2	0.72
Compound4	1.66	**1.28**	1.04	3.99	6.67	6.75	**1**	1	1.28
Compound5	0.55	**0.85**	0.64	1.09	1.23	1.97	**2**	4	1.23

[a] Code1 is the coded information of the best predicting model for the compound.
[b] Code2 is the predicted coded information by a random forest.

Model evaluation – Comparative prediction accuracy

The prediction accuracy of each Classification tree was compared to the prediction accuracy for each existing algorithm within its group (Table 3-4). This means that, using inputs required by the Classification Tree with all others estimated based on Table 3-3, the prediction accuracy of the Classification Tree was compared to the prediction accuracy of each algorithm in the group. Prediction accuracy was based on a comparison of the predicted and observed K_ps for each algorithm. To assess the overall precision of each algorithm, the root mean squared error (RMSE) was calculated (Eqn. 3-20) as well as the overall percentage within k-fold deviation (k=1.25, 1.5, 2, 3). Tissue specific RMSE was also calculated for comparison of precision of the models with respect to the tissue. As a measure of bias, the average fold error (AFE) was calculated for each Classification tree (Eqn. 3-18). The AFE indicates an under-prediction (AFE < 1) or an over-prediction (AFE > 1) compared to the observed values. The absolute average fold error (AAFE) quantifies the overall magnitude of the deviation between the predicted and the observed K_p values (Eqn 3-1). Second, using the predicted values from previously published algorithms (e.g. Jansson *et al.*[5], Rodgers *et al.* [8,9]) the same procedure (i.e. % within k-fold error, AFE, AAFE, global and tissue specific RMSE calculations) was conducted. The accuracy of prediction for each Classification tree was compared to each of the previously published algorithms within its group to assess if any one previously published algorithm performed better than the Classification tree.

3.4 Results

Dataset

The dataset was comprised of a total of 122 compounds with 852 K_ps in 11 tissues (Appendix 7 and 8). The physicochemical properties and *in vivo* properties were gathered from the literature. The dataset consisted of 29 acids, 70 bases (63 moderate to strong bases with pKa ≥ 7.4 and 7 weak bases with pKa ≤ 7.4), 12 neutrals, and 11 zwitterions (Figure 3-2).

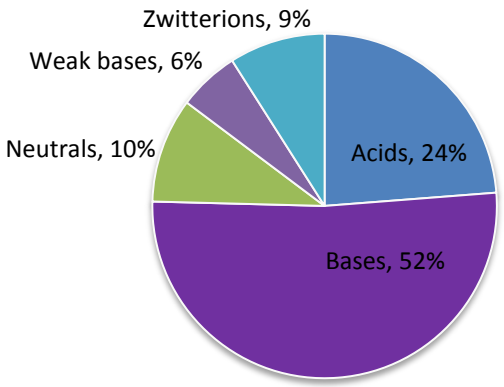


Figure 3-2. Proportion of molecular species of compounds in the total dataset

K_p calculations according to the previously published algorithms

Predicted K_ps as published by existing algorithms were compared to K_ps predicted using experimental input data and estimation equations for input parameters not required for Classification tree use. The K_p predictions deviated from the original published predictions (Table 3-7, Table 3-8, Table 3-9); however, the mean fold error per tissue was comparable to that in the original publications.

Table 3-7. Comparison of predicted K_ps from Rodgers *et al.* [8,9] vs. those predicted using experimental/estimated input parameters.

		Adipose	Bone	Brain	Gut	Heart	Kidney	Liver	Lung	Muscle	Skin	Spleen
K_p predictions from Rodgers *et al.* [8]	MFE	1.41	4.87	1.62	0.90	1.35	0.93	1.44	0.64	1.57	2.00	1.16
	AFE	1.13	1.27	0.80	0.55	1.17	0.75	0.92	0.44	1.32	1.90	0.83
	AAFE	1.82	2.33	2.62	2.50	1.48	1.76	1.97	2.56	1.62	1.90	2.06
	RMSE	0.33	0.57	0.50	0.53	0.23	0.32	0.36	0.48	0.27	0.32	0.38
K_p prediction using the experimental/estimated inputs	MFE	1.48	7.09	1.48	0.76	1.52	0.96	1.53	0.71	1.62	1.91	0.97
	AFE	1.14	1.17	0.70	0.51	1.22	0.82	0.95	0.43	1.32	1.79	0.78
	AAFE	1.85	2.66	2.89	2.48	1.71	1.63	2.04	2.72	1.62	1.82	1.92
	RMSE	0.34	0.66	0.57	0.48	0.29	0.26	0.40	0.53	0.28	0.30	0.30
		Adipose	Bone	Brain	Gut	Heart	Kidney	Liver	Lung	Muscle	Skin	Spleen
K_p predictions from Rodgers & Rowland. [9]	MFE	1.28	0.67	3.16	1.36	1.06	0.57	0.65	1.40	1.02	2.00	1.23
	AFE	0.97	0.52	2.17	1.04	0.89	0.42	0.45	1.16	0.88	1.69	0.96
	AAFE	1.91	2.05	2.31	1.82	1.68	2.57	2.64	1.67	1.51	1.79	1.72
	RMSE	0.39	0.47	0.47	0.34	0.27	0.53	0.53	0.27	0.24	0.33	0.31
K_p prediction using the experimental/estimated	MFE	1.10	0.70	3.26	1.37	1.24	0.61	0.68	1.56	0.95	2.08	1.18
	AFE	0.58	0.58	1.87	0.97	0.86	0.39	0.37	1.11	0.77	1.67	0.85

inputs	AAFE	2.82	1.82	2.29	1.98	1.78	2.89	3.22	1.74	1.63	1.91	1.89
	RMSE	0.66	0.37	0.49	0.39	0.33	0.59	0.63	0.32	0.29	0.37	0.34

Table 3-8. Comparison of predicted K_ps from Jansson *et al*. [5] vs. those predicted using experimental/estimated input parameters.

	Metrics	Adipose	Bone	Brain	Gut	Heart	Kidney	Liver	Lung	Muscle	Skin
K_p predictions from Jansson *et al*. [5]	MFE	1.61	0.77	1.51	1.48	1.20	2.03	4.77	1.90	1.04	1.14
	AFE	0.85	0.74	1.15	1.23	0.95	1.41	1.88	1.47	0.95	1.00
	AAFE	2.16	1.46	1.80	1.66	1.77	2.27	2.46	1.91	1.48	1.47
	RMSE	0.43	0.17	0.33	0.30	0.38	0.44	0.57	0.36	0.20	0.21
K_p predictions using observed Muscle K_p	MFE	2.40	0.78	1.79	1.61	1.13	1.91	3.12	1.46		1.14
	AFE	1.05	0.74	1.15	1.19	0.96	1.34	1.39	1.20		0.99
	AAFE	3.05	1.43	2.30	1.92	1.49	2.14	2.29	1.74		1.51
	RMSE	0.57	0.20	0.48	0.35	0.30	0.41	0.49	0.31		0.24
K_p predictions using estimated Muscle K_p	MFE	2.34	0.77	1.52	1.48	1.13	2.03	4.77	1.76	1.04	1.14
	AFE	1.13	0.74	1.08	1.23	0.93	1.41	1.88	1.36	0.95	1.00
	AAFE	2.67	1.46	1.84	1.66	1.75	2.27	2.46	1.78	1.48	1.47
	RMSE	0.53	0.17	0.35	0.30	0.35	0.44	0.57	0.34	0.20	0.21

Table 3-9. Comparison of predicted K_ps from Schmitt [6] vs. those predicted using experimental/estimated input parameters.

		Adipose	Bone	Brain	Gut	Heart	Kidney	Liver	Lung	Muscle	Skin	Spleen
K_p predictions from Schmitt [6]	MFE	9.06	3.43	12.59	2.70	4.09	1.27	1.91	1.20	1.78	4.89	0.92
	AFE	4.45	1.30	7.18	1.35	2.52	0.68	0.80	0.77	1.23	2.82	0.78
	AAFE	4.63	2.73	7.18	2.65	2.90	2.26	2.42	2.03	1.84	3.02	1.64
	RMSE	0.83	0.56	0.97	0.53	0.57	0.45	0.52	0.40	0.35	0.62	0.30
K_p predictions using experimental/estimated inputs	MFE	9.25	5.49	13.25	2.10	3.94	0.99	1.65	1.47	1.49	4.27	1.20
	AFE	4.81	1.01	6.63	1.23	2.73	0.65	0.82	0.75	1.16	2.92	0.93
	AAFE	4.89	2.75	7.36	2.46	3.09	2.01	2.36	2.34	1.83	3.03	2.05
	RMSE	0.84	0.62	0.97	0.48	0.58	0.37	0.48	0.49	0.32	0.57	0.34

For K_p calculation according to Rodgers *et al.* [8], the prediction accuracy based on the use of the previously published estimation equation for B:P (Eqn. 3-11) and the developed regression equation (Eqn. 3-12) was compared. The use of the developed regression equation resulted in a more accurate prediction in K_ps with lower tissue specific RMSE values (Table 3-10). As a result, the developed regression equation (Eqn. 3-12) was used in all subsequent calculations.

Table 3-10. Comparison of K_p prediction accuracy based on the Rogers et al. [8] algorithm using either the Paixao *et al.* [74] B:P estimation equation or the regression equation developed in this study.

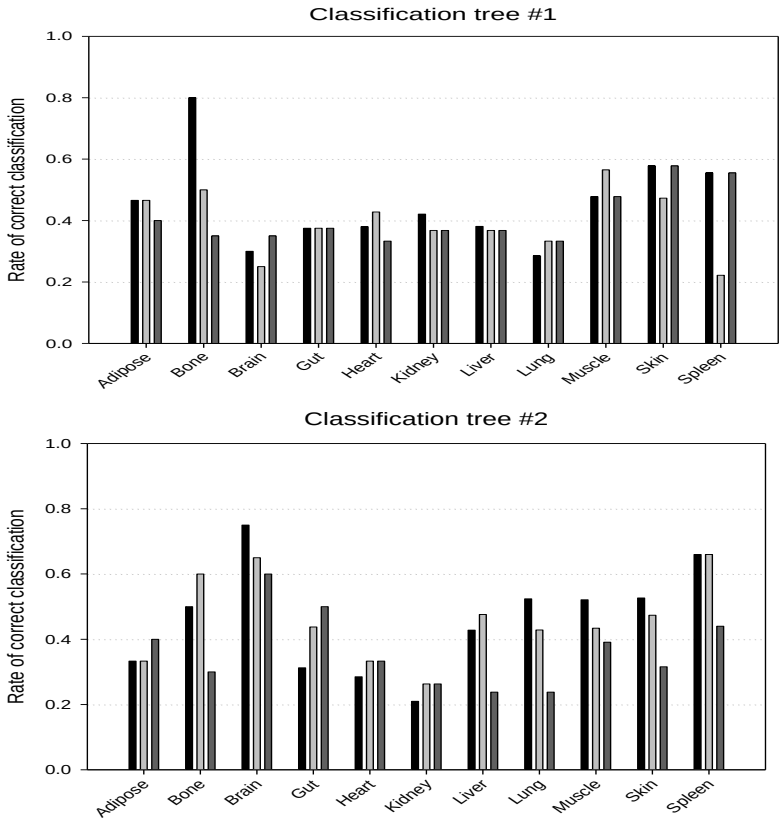
B:P estimation Method	RMSE										
	Adipose	Bone	Brain	Gut	Heart	Kidney	Liver	Lung	Muscle	Skin	Spleen
Paixao *et al.* [74] (Eqn. 3-12)	0.51	0.84	0.78	0.87	0.68	0.75	0.74	1.07	0.58	0.39	0.96
Regression equation (Eqn. 3-11)	0.34	0.66	0.56	0.48	0.29	0.27	0.41	0.54	0.29	0.31	0.31

With the use of estimated input parameters (e.g. B:P, LogKvo:w), the K_ps calculated using the algorithm of Rodgers *et al.* [8,9] resulted in a under-prediction when compared to K_ps calculated by the author with the experimentally determined parameters (Table 3-7). For Jansson *et al.* [5] and Schmitt [6], with the use of estimated input parameters (Eqn. 3-6, 3-7, 3-8, 3-12), the K_ps calculated using each algorithm were in agreement with the K_ps obtained by both Jansson *et al.* [5] and Schmitt [6] (Table 3-8, Table 3-9, respectively).

Investigation of various classification methods

Decision trees were developed for 11 tissues as these contained a sufficient number of data points for development (Table 3-11). Among several classification methods (i.e. random forest, bagging, recursive partitioning), classification performance was explored using the same set of

the data. Based on the rate of correct classification, random forest was superior to others with the highest correct classification rates in the majority of tissues among each set (Figure 3-3). However, the magnitude and standard deviation are similar among the different methods. Thus, random forest was deemed to classify the most accurate K_p prediction model based on the physicochemical space of compounds.



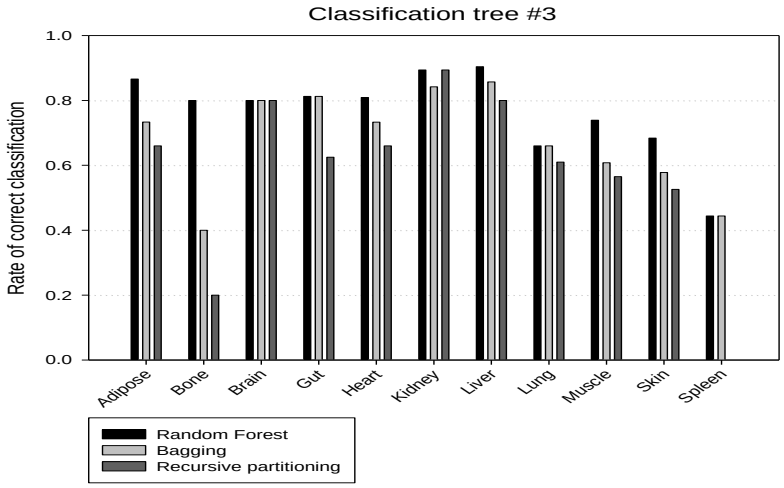


Figure 3-3. Rates of correct classification of various classifier algorithms with respect to a tissue**.**

Descriptive statistics of K_p algorithm performance based on the chemical properties

Using the dataset that consists of 122 compounds, K_ps were calculated according to the published algorithms. The best prediction algorithm for each compound-tissue combination was assessed. This information was stratified by the compound's acid-base-neutral properties (Figure 3-4, left), and LogP values (Figure 3-4, right). For example, for basic compounds, 27% of K_ps were best predicted by Yun and Edginton [19]. For compounds with a LogP value between -3 and 1, 27% were more accurately predicted by Jansson *et al.* [5] (Figure 3-4, right).

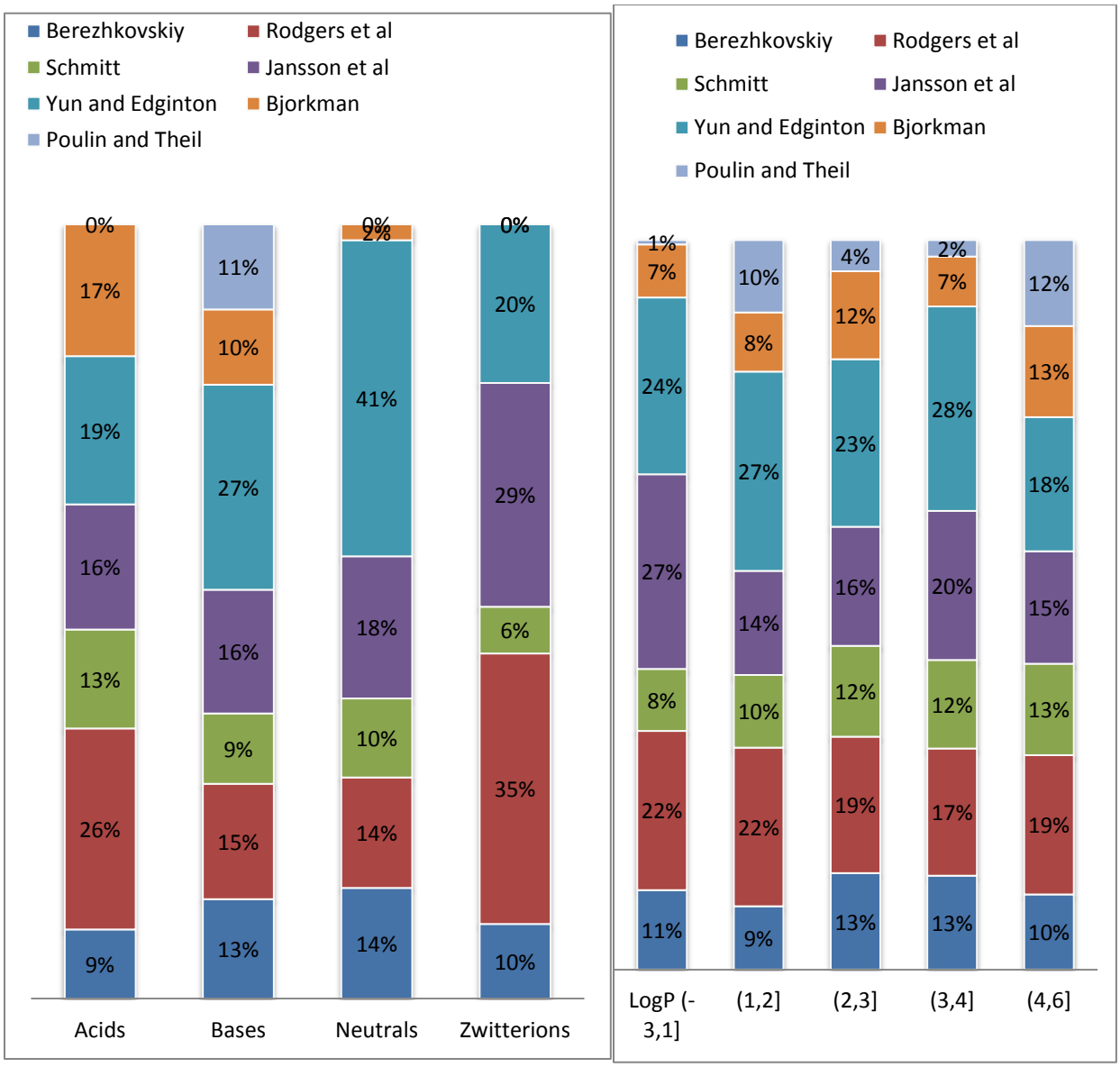


Figure 3-4. Schematics of the best prediction algorithms based on molecular species (left), and lipophilicity (right) in the total dataset (n=122 compounds)

Construction of predictive random forest models: Classification tree #1, # 2 and # 3

Three Classification trees were developed using the random forest method. The number of samples and the chosen m_{try} are listed in (Table 3-11). The classification performance of each classification tree was indicated by the rate of correct classification. Classification trees resulted in a greater rate of correct classification than random permutation rates of 1/6, 1/5, 1/3, based on the probability of a correct classification when there are n categories, (1/n). The prediction accuracy for each Classification tree was indicated by the percentage of predicted values within 2 fold of the observed K_ps for each tissue. Based on Table 3-11, a high rate of correct classification did not always improve K_p prediction accuracy (i.e. percentage within 2 fold of deviation from the observed K_ps), especially in Classification tree #3. The rate of correct classification for Classification tree #1 and #2 was relatively lower than that of Classification tree #3. This was because Classification tree #3 had only two or three algorithms to classify whereas Classification tree #1 had 5 to 6 and Classification tree #2 had 4 to 5 (Table 3-4).

Table 3-11. Summary of random forest parameter and classification performance.

	Classification tree #1				Classification tree #2				Classification tree #3			
	n	m_{try}	Rate of correct classification	% within 2 fold error	n	m_{try}	Rate of correct classification	% within 2 fold error	N	m_{try}	Rate of correct classification	% within 2 fold error
Adipose	66	5	0.359	51.6%	65	2	0.384	54.6%	69	4	0.638	60.0%
Bone	41	5	0.561	73.2%	41	5	0.561	75.6%	42	2	0.643	50.0%
Brain	78	5	0.385	56.4%	76	5	0.395	51.3%	90	4	0.644	47.8%
Gut	68	5	0.368	72.1%	65	5	0.446	80.0%	68	4	0.618	60.3%
Heart	91	5	0.452	83.3%	83	5	0.446	80.7%	96	4	0.563	60.4%
Kidney	89	5	0.341	73.9%	86	5	0.386	69.8%	94	4	0.684	55.3%
Liver	84	5	0.243	64.2%	84	5	0.429	63.1%	88	4	0.693	51.1%
Lung	93	5	0.312	67.8%	85	5	0.365	64.7%	95	2	0.589	56.8%
Muscle	108	5	0.630	78.7%	93	5	0.355	79.6%	108	4	0.667	80.5%
Skin	64	5	0.328	77.4%	61	5	0.393	77.1%	64	2	0.719	71.9%
Spleen	36	5	0.583	61.1%	33	2	0.424	63.6%	36	4	0.528	58.3%

Comparative assessment of K_p *prediction accuracy of Classification trees and published equations*

-Comparison of prediction accuracy of classification tree # 1 and published equations

In order to compare the predictive performance of the published algorithms [4-9,12] and Classification tree #1, the tissue AFE, AAFE, and RMSE were calculated using the same dataset (Appendix 7, Appendix 8). A plot of percentage within k- fold deviation from observed values showed that predictions based on Classification tree #1 performed well with 25.6%, 49.7% and 68.8% falling within 1.25, 1.5 and 2 fold deviation from the observed K_p values, respectively (Figure 3-5). Global RMSEs of algorithms in Group 1 indicated that the K_p prediction errors are similar for Jansson *et al.* [5], Rodgers *et al.* [8,9], and Classification tree #1 with values 0.43, 0.51 and 0.49 (Table 3-12). However, Rodgers *et al.* [8,9] and Classification tree #1 tended to under-predict K_p with AFE values of 0.89, and 0.94, respectively. The under-prediction in K_ps of Rodgers *et al.* [8,9] was observed in bone, kidneys and liver. Jansson *et al.* [5] had the smallest RMSE values of 0.43 but appeared to over-predict K_p with the AFE of 1.27 (Figure 3-6, Table **3-12**). The over prediction of K_ps by Jansson *et al.* [5] was observed in kidneys, liver and adipose tissue. The overall bias of deviation between the observed K_ps and those estimated using Classification tree #1 was the smallest among Group 1 with the AFE value of 0.94 (Table 3-12). This is further supported by the tissue specific box whisker plot, where the boxes for Classification tree #1 are small, centered around zero, and not showing evidence of serious under- or over- prediction. Tissue specific RMSEs showed that the K_p prediction of Jansson *et al.* [5] resulted in the smallest error for 6 out of 11 tissues in Group 1 (Table 3-13). It was observed that Berezhkovskiy [12], Schmitt [6] and Bjorkman's models [4] tended to over-predict the K_ps with an AFE value larger than 1 (Table 3-12). On the other hand, Rodgers *et al.* [8,9] and Poulin and Theil's [7] models tended to under-predict the K_ps with an AFE value less than 1.

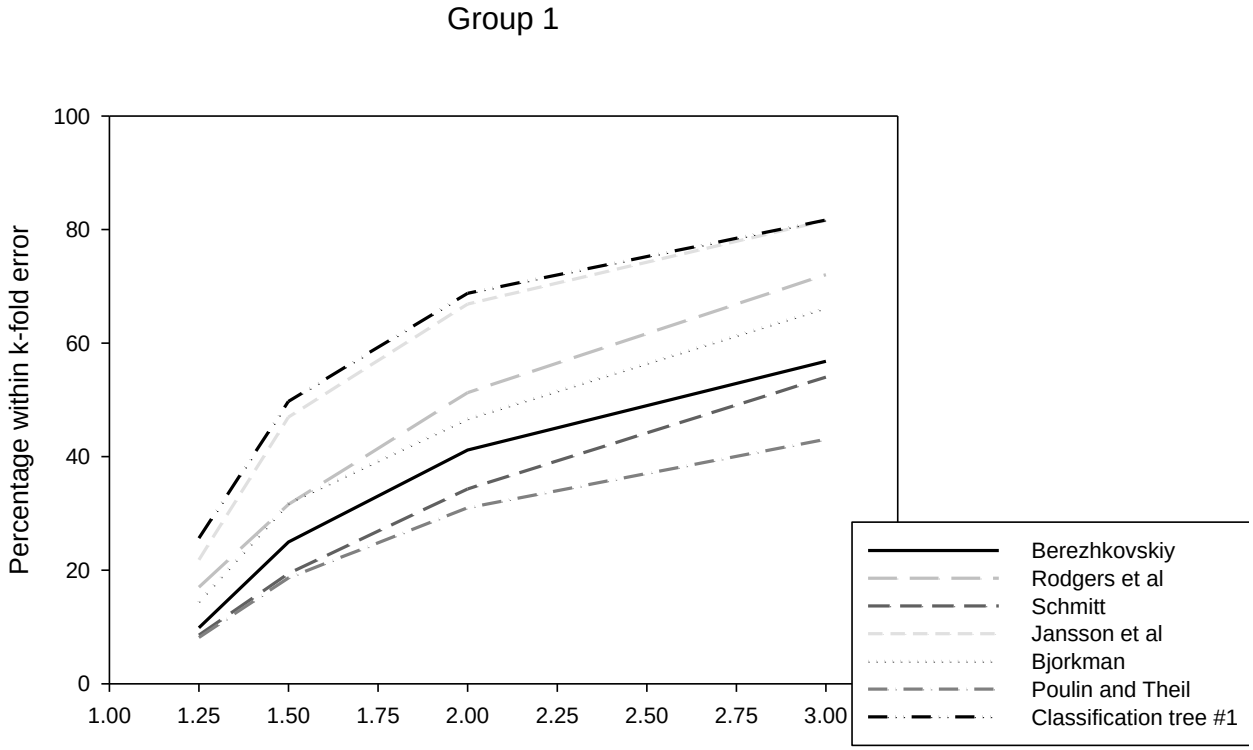


Figure 3-5. Percentages within k fold error. X-acids represents folds, y-axis represent the percentage within k fold error of deviation in Group 1.

Table 3-12. Summary of overall predictive performance for Group 1.

	Berezhkovskiy [12]	Rodgers *et al.* [8,9]	Schmitt [6]	Jansson *et al.* [5]	Bjorkman [4]	Poulin and Theil [7]	Classification tree #1
AFE	1.14	0.89	1.37	1.27	1.52	0.16	0.94
AAFE	3.21	2.34	3.36	1.98	2.81	8.34	2.00
RMSE	0.67	0.51	0.66	0.43	0.62	1.25	0.49

Table 3-13. Summary of tissue specific RMSE of different algorithms in Group 1.

	Berezhkovskiy [12]	Rodgers *et al.* [8,9]	Schmitt [6]	Jansson *et al.* [5]	Bjorkman [4]	Poulin and Theil [7]	Classification tree #1
Adipose	0.79	0.47	0.85	0.75	1.20	1.72	0.77
Bone	0.60	0.55	0.65	0.49	0.64	1.62	0.44
Brain	0.84	0.58	1.02	0.43	.62	1.39	0.75
Gut	0.59	0.39	0.50	0.31	0.45	0.72	0.44
Heart	0.50	0.34	0.63	0.26	0.49	1.08	0.26
Kidney		0.64	0.54	0.33	0.47	0.93	0.38
Liver		0.65	0.59	0.51	0.54	1.25	0.54
Lung	0.76	0.50	0.57	0.34	0.55	1.42	0.37
Skin	0.45	0.41	0.56	0.23	0.40	1.00	0.32
Spleen	0.64	0.34	0.51			1.02	0.34

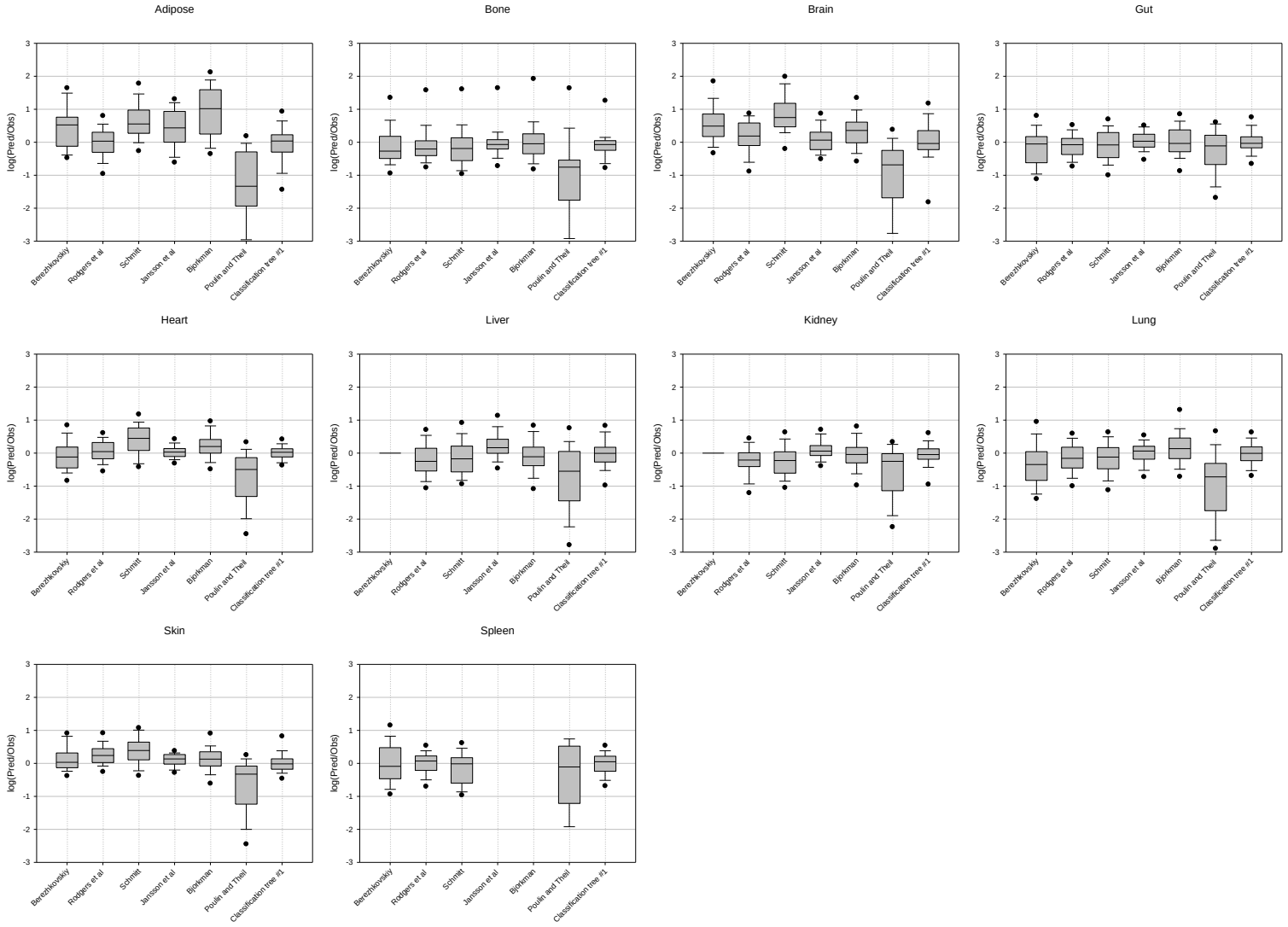


Figure 3-6. Box and Whisker plot of the logarithm of the ratio between the predicted and observed K_p values of predicted K_ps from published equations in Group 1 and random forest (Classification tree #1). The boxes represent the median (line) and the 25^{th} and 75th percentiles; the bars represent the 10^{th} and 90^{th}. The dots are the 5^{th} and 95^{th} percentiles.

-Comparison of prediction accuracy of classification tree# 2 and published equations

For comparison of the predictive performance of the published algorithms [4-9,12,19] and Classification tree #2, the tissue AFE, AAFE, and RMSE were calculated. Both Classification tree #2 and Yun and Edginton [19] resulted in more accurate K_p predictions with higher percentages within k- fold deviation from observed K_ps (k = 1.25 to 3) compared to other algorithms. The prediction performances of both Classification tree #2 and Yun and Edginton's algorithm [19] were very similar with almost the same AFE, AAFE, global RMSE and tissue specific RMSE values (Table 3-14, Table 3-15).

Favorable K_p predictive performance of both Classification tree #2 and Yun and Edginton [19] algorithms was further reinforced by their AFE values which were closest to 1, and their small AAFE values less than 2. The plot of percentage within k- fold deviation from observed values showed that Classification tree #2 based K_p prediction performed well with 31.9% and 50.4% falling within 1.25 and 1.5 fold deviation from the observed K_p values, respectively (Figure 3-7).

In 6 out of 11 tissues, Yun and Edginton algorithm [19] resulted in the smallest error associated with K_p estimates (Table 3-15). Jansson *et al.* [5] showed an over-prediction in K_ps that was mainly due to the over-prediction in the adipose and liver K_ps (Figure 3-8). Schmitt's algorithm [6] tended to over-predict K_ps with an AFE of 1.28 and was less accurate with an AAFE of 3.20 (Table 3-14). An over-prediction in K_ps by Schmitt [6] was observed in adipose, brain, heart and skin (Figure 3-8). Although Berezhkovskiy's [12] algorithm resulted in an AFE value close to 1 (1.02), its AAFE value was 2.92. This implies that K_p predictions were less accurate and there were both under and over-predictions in the K_ps. The box whisker plot showed that there was over-prediction in the brain and adipose tissue K_ps and an under-prediction in gut and lung K_ps.

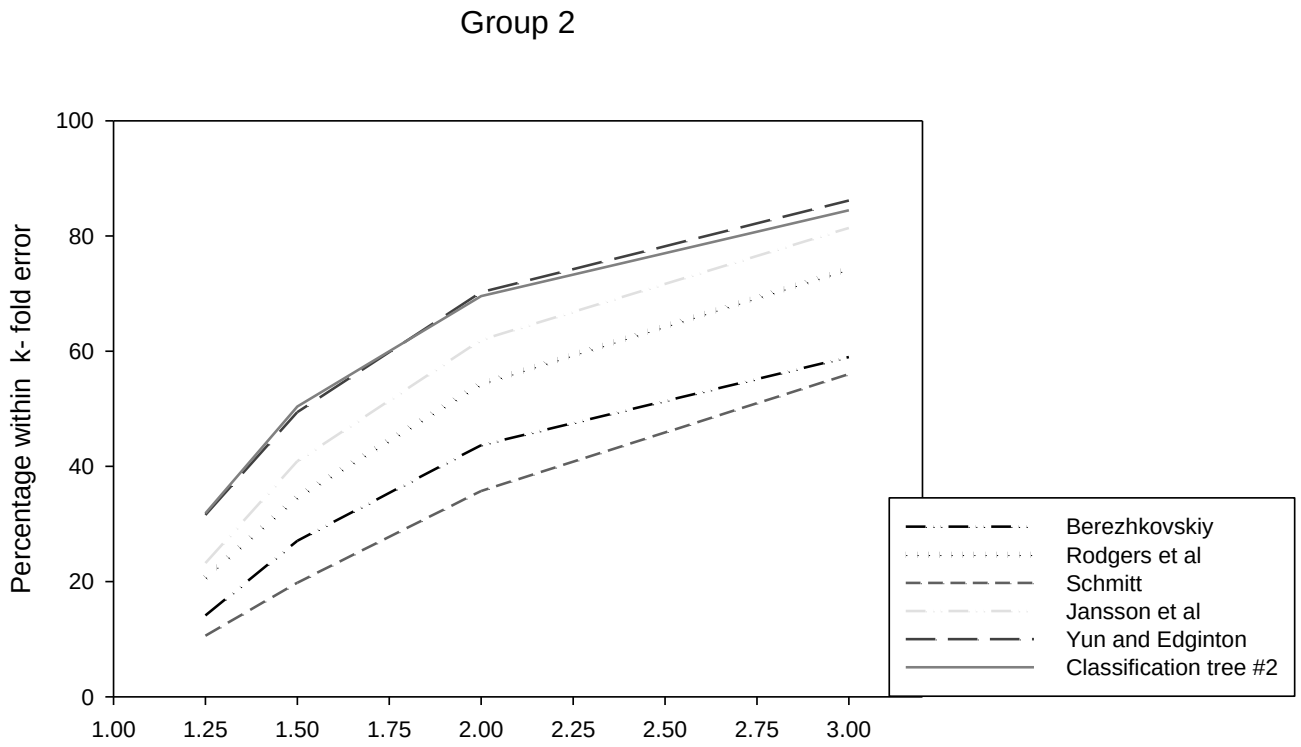


Figure 3-7. Percentage within k-fold error. X-axis represents folds, y-axis represent the percentage within k fold error of deviation in Group 2.

Table 3-14. Summary of overall predictive performance for Group 2.

Group 2	Berezhkovskiy [12]	Rodgers *et al.* [8,9]	Schmitt [6]	Jansson *et al.* [5]	Yun and Edginton [19]	Classification tree #2
AFE	1.02	0.93	1.28	1.21	1.01	1.03
AAFE	2.92	2.20	3.20	2.06	1.78	1.82
RMSE	0.60	0.45	0.64	0.45	0.36	0.37

AFE: average fold error, AAFE: absolute average fold error, RMSE: root mean square error

Table 3-15. Summary of tissue specific RMSE of different algorithms in Group 2.

	Berezhkovskiy [12]	Rodgers *et al.* [8,9]	Schmitt [6]	Jansson *et al.* [5]	Yun and Edginton [19]	Classification tree #2
Adipose	0.78	0.48	0.85	0.78	0.45	0.50
Bone	0.60	0.55	0.65	0.51	0.52	0.43
Brain	0.73	0.57	0.97	0.47	0.50	0.48
Gut	0.60	0.38	0.49	0.28	0.25	0.25
Heart	0.46	0.34	0.62	0.42	0.25	0.31
Kidney		0.49	0.54	0.35	0.37	0.36
Liver		0.56	0.58	0.50	0.38	0.43
Lung	0.73	0.47	0.59	0.41	0.32	0.36
Muscle	0.48	0.29	0.46	0.33	0.28	0.28
Skin	0.37	0.39	0.54	0.28	0.28	0.26
Spleen	0.53	0.33	0.52		0.26	0.32

RMSE: root mean square error

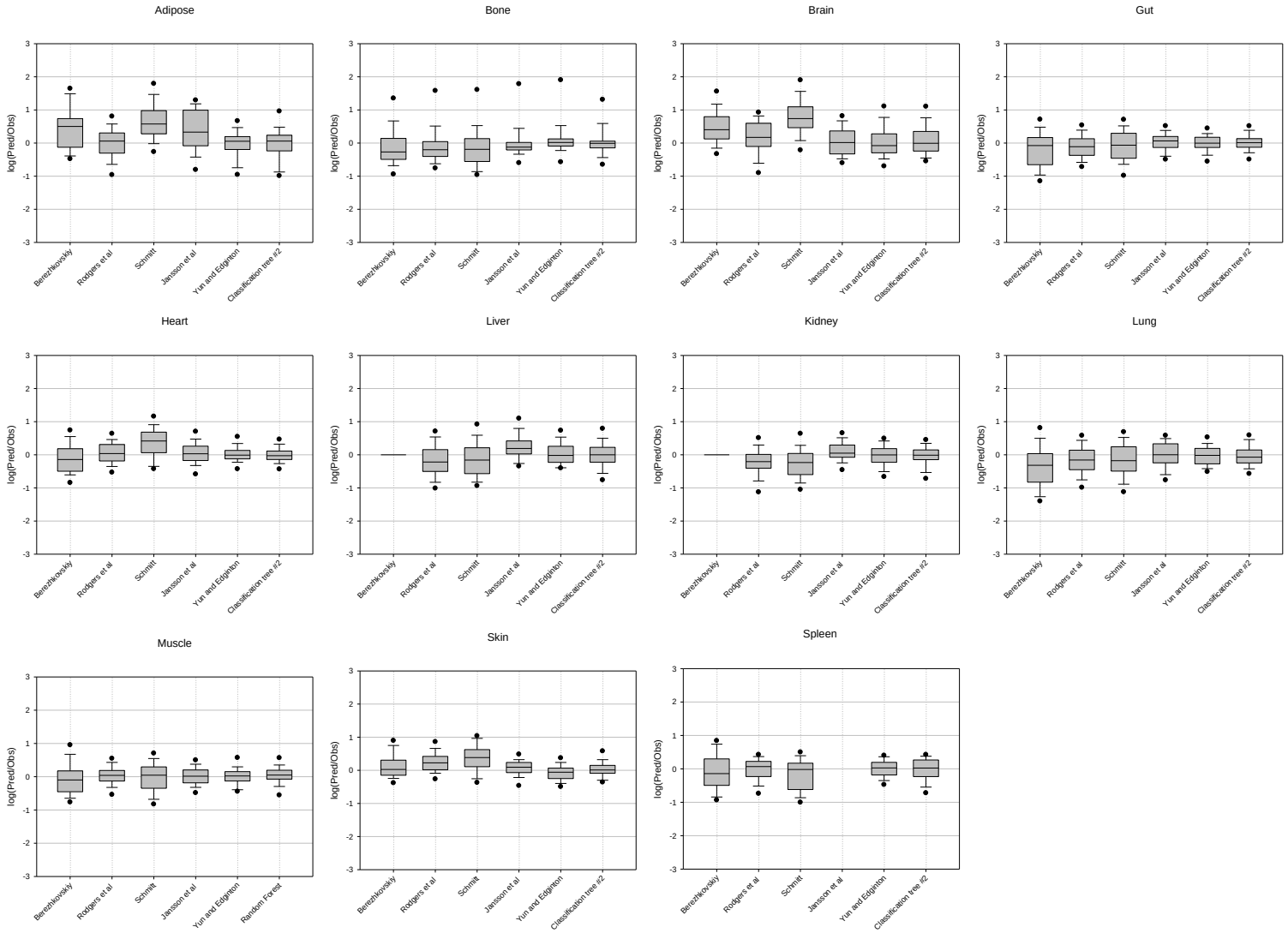


Figure 3-8. Box and Whisker plot of the logarithm of the ratio between the predicted and observed K_p values of predicted K_ps from published equations in Group 2 and random forest (Classification tree #2). The boxes represent the median (line) and the 25th and 75th percentiles; the bars represent the 10th and 90th. The dots are the 5th and 95th percentiles.

-Comparison of prediction accuracy of classification tree# 3 and published equations

For comparison of the predictive performance of the published algorithms [6,8,9,12] and Classification tree #3, the tissue AFE, AAFE, and RMSE were calculated. Classification tree #3 resulted in accurate predictions in Group 3 with the highest percentages within k-fold deviation from observed K_ps (Figure 3-9), the smallest global RMSE of 0.45, AFE of 0.95 and the smallest AAFE of 2.14. In 9 out of 11 tissues, Classification tree #3 resulted in the smallest tissue specific RMSEs. The Berezhkovskiy [12] and Schmitt [6] algorithms were less accurate with an AAFE larger than 3 and both had a tendency to over-predict the K_ps with an AFE value larger than 1 (Table 3-16). Rodgers *et al*. [8,9] under-predicted the K_ps with an AFE of 0.91. An under-prediction in the K_ps by Rodgers *et al*. was observed in bone, kidneys, liver and lungs (Figure 3-10). The global RMSE, AFE, and AAFE values for Classification tree #1, #2 and Classification tree #3 were comparable. However, in the case of Classification tree #3, the percentage within k-fold deviation from observed K_ps was lower than Classification tree #1 and #2.

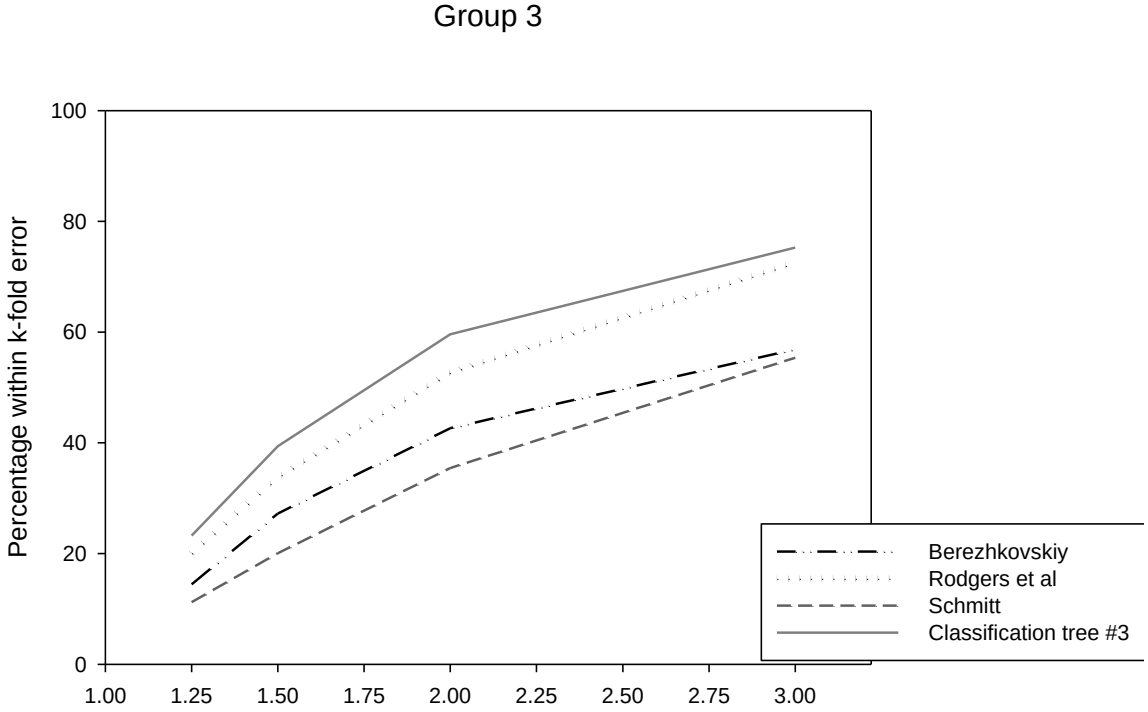


Figure 3-9. Percentage within k-fold error. X-axis represents folds, y-axis represent the percentage within k-fold error of deviation in Group 3.

Table 3-16. Summary of overall predictive performance for Group 3.

Group 3	Berezhkovskiy [12]	Rodgers *et al.* [8,9]	Schmitt [6]	Classification tree #3
AFE	1.16	0.91	1.37	0.95
AAFE	3.18	2.33	3.27	2.14
RMSE	0.66	0.52	0.65	0.45

Table 3-17. Summary of tissue specific RMSE of different algorithms in Group 3.

	Berezhkovskiy [12]	Rodgers *et al.* [8,9]	Schmitt [6]	Classification tree #3
Adipose	0.82	0.47	0.84	0.45
Bone	0.59	0.54	0.65	0.54
Brain	0.85	0.61	1.00	0.58
Gut	0.59	0.39	0.50	0.36
Heart	0.49	0.36	0.65	0.37
Kidney		0.64	0.54	0.45
Liver		0.71	0.57	0.53
Lung	0.75	0.50	0.57	0.46
Muscle	0.51	0.37	0.47	0.30
Skin	0.45	0.41	0.56	0.35
Spleen	0.64	0.34	0.51	0.35

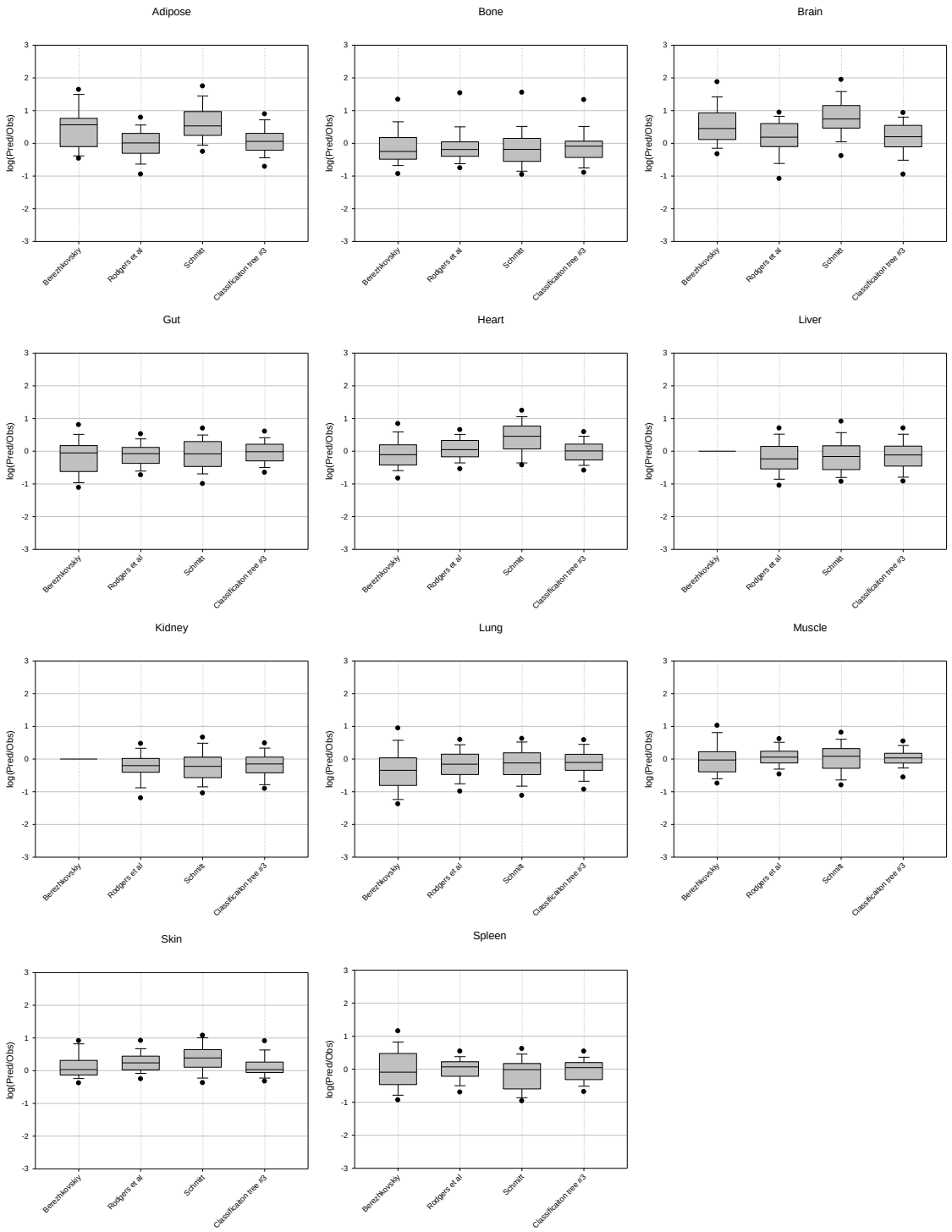


Figure 3-10. Box and Whisker plot of the logarithm of the ratio between the predicted and observed K_p values of predicted K_ps from published equations and random forest (Classification tree #3). The boxes represent the median (line) and the 25th and 75th percentiles; the bars represent the 10th and 90th. The dots are the 5th and 95th percentiles.

3.5 Discussion

K_p predictions with estimated input parameters

One of the objectives of this study was to develop a tool to provide K_p prediction when only a limited number of parameters are available. Many algorithms require input parameters that are not readily available to researchers such as muscle K_p or B:P. As a result, Classification trees were built using experimental input parameters that are readily available while estimating those that are not considered routinely derived. To assess the use of estimation methods for generally unavailable input parameters, a comparison of predicted K_ps from published algorithms were compared to the predicted K_ps using readily available experimental parameters and the estimated input parameters.

In the calculation of Rodgers *et al.* [8,9], it was observed that the use of experimentally determined inputs such as B:P and LogKvo:w resulted in more accurate K_p predictions with lower tissue specific RMSEs when compared to K_ps calculated using estimated inputs (Table 3-7, Table 3-8, Table 3-9). In Rodgers *et al.* [8], the blood cell to plasma water concentration ratio (K_puBC) is one of the parameters that is not directly measured but is estimated using a standard equation (Eqn. 3-10). This equation is a function of an experimentally determined B:P [73]. Therefore, the prediction of K_ps according to Rodger *et al.* [8] is sensitive to the accuracy of the B:P measurement. Instead of using an experimentally determined B:P, Small et al. [76] introduced an alternative method that directly measures K_puBC using surface plasmon resonance (SPR). It was discovered that the use of the SPR approach resulted in a more accurate prediction of K_pu and therefore Vss [76]. This demonstrates that the more accurate the input, the more accurate the predictions. Availability of either experimentally determined B:P or K_puBC is likely to lead to a more accurate K_p prediction using the algorithm of Rodgers et al. [8,9]. In reality however, these parameters are not often available. In order to overcome this problem, a B:P estimation equation (Eqn. 3-12) was generated in this study. This equation was used and replaced the previously published

estimation approach (Eqn 3-11 [74]) as the regression equation produced more accurate $K_p$us (Table 3-10). The use of this regression equation may bring uncertainty to our model. However, the use of this equation in K_p calculations using the Rodgers *et al.* [8,9] algorithm resulted in $K_p$us that were comparable, although not superior to, $K_p$us calculated using experimentally determined B:P. The accuracy metrics such as tissue specific RMSEs and AFEs were comparable (Table 3-7).

In the calculation of Jansson *et al.*'s algorithm [5], the use of an experimentally determined muscle K_p resulted in more accurate predictions in heart, kidney, liver and lung when compared to the prediction accuracy of Jansson *et al.* [5] that used a muscle K_p that was estimated from Vss (Table 3-8). As a result, Jansson *et al.*'s algorithm [5] was selected as the best predicting algorithm in Classification tree #1, which used muscle K_p as an input, more often than in Classification tree #2, which used Vss as an input. Overall, based on similar bias and precision estimates, K_p predictions with the estimated input parameters were deemed sufficiently agreeable to K_p predictions from Rodgers *et al.* [8,9], Jansson *et al.* [5], and Schmitt [6].

Construction of tissue specific Classification trees #1, #2, and #3

Because compound distribution is the interplay between compound specific properties (pKa, LogP, and fup) and physiologic factors such as tissue composition information (e.g. concentration of acidic phospholipids), K_p prediction equations should be able to describe the compound distribution process affected by both the physicochemical properties of a compound and the tissue specific physiologic factors. Those factors should be well formulated to yield a sufficient prediction. Failure to take into account one of the above aspects could result in K_p predictions deviating from the true value.

The predictive performance of a K_p algorithm may be tissue-dependent. One algorithm may have more predictive power for a particular tissue than an alternative algorithm. Furthermore, accurate K_p

prediction in some tissues is more difficult than others. For example, K_p predictions in lung, adipose, and liver are difficult due to the enhanced probability of ion trapping, the large distribution of lipophilic compounds into adipose tissue with a relatively large inter-laboratory measurement error on LogP (explained in the discussion of Chapter 2) and the role of extraction in K_p estimates. In order to address inter-tissue variability, a Classification tree was created for 11 tissues. For each tissue, Classification trees #1, #2 and #3 were constructed that were dependent upon user supplied input parameters (i.e. LogP, pKa, fup, Vss, and muscle K_p) as well as estimated input parameters that were required but not deemed readily available.

Comparison between classification methods

In the generation of Classification trees, the classification performances of the three different classification algorithms (i.e. recursive partitioning, bagging, and random forest methods) were investigated. The algorithm with the highest correct classification ratio was selected for this study. The three classification methods differ in their methodologies. One of the disadvantages of using a single classification tree derived from the recursive partitioning method is that it can be sensitive to the modifications in the training set when compared to a collection of classification trees [61]. The single classification tree is unstable due to the numerous potential variables that can lead to a reduction in impurity when a split is chosen. In other words, depending on the dataset different splitting criteria can be chosen for a node resulting in a different classification. In order to overcome the instability of the single classification tree, ensemble methods (i.e. random forest and bagging) are used.

In the bagging and random forest method, because of the random variation of each bootstrap sample drawn from the training data, various classification trees with different splitting criteria were generated. By combining the classifications from the trees, there is an increase in the correct classification ratio

when bagging or random forest is used. However, the easy interpretability of the single classification tree (e.g. Figure 3-1) is not available as an output of the ensemble methods.

Both random forest and bagging are similar in the use of the same recursive partitioning principle when growing a collection of trees. Random forest and bagging methods are different in that, with random forest the splitting criteria is chosen from the m_{try} variable. In the bagging model, the splitting criteria is chosen from all of the M number of variables [69]. Random forest grew 500 trees whereas, bagging grew 25 trees by default in this analysis. Therefore, by optimizing m_{try} in the random forest, more various trees can be grown from the bootstrap subsets than with the bagging method.

It was observed that, in most tissues, Classification trees #1, #2, and #3 were optimized with m_{try} values close to the maximum number of input variables (e.g. for group 3, M=4: LogP, DOI, fup, and class) (Table 3-4). In most cases, the number of variables at each node were the same with m_{try}=M in both the random forest and bagging methods. Whereas for bagging, m_{try} was always set to be M (i.e. m_{try}=M). Random forest grew trees with a different m_{try} and among the possible m_{try}, the optimal m_{try} was found by selecting m_{try} that resulted in the smallest cross validation error. The large number of trees and optimized m_{try} of random forest led to a more precise classification in this study. Among the classification methods, random forest was selected for this study due to the higher rate of correct classifications in most tissues (Figure 3-3).

Inherent factors in K_p *prediction via a Classification tree*

K_p prediction via a Classification tree depends on two important factors. The first factor is the accuracy of each K_p prediction algorithm in each group (e.g. Rodgers *et al.* [8,9], Jansson *et al.* [5]), and the second factor is the classification performance of a classifier (i.e. a random forest). Although poor prediction of the K_ps and/or poor classification by a classifier can lead to an undesirable outcome, there is no clear

relationship between the accuracy of a K_p prediction method and the classification performance. The rate of correct classification did not always result in the lowest RMSE even though the best performing algorithm (e.g. Yun and Edginton [19]) for a certain compound was correctly predicted. This is because the predicted K_p from an algorithm that was classified by the random forest can largely deviate from the corresponding observed K_p (Table 3-11). Thus, the interplay of these two factors should be taken into consideration in the interpretation of the K_p prediction via the Classification trees #1, #2 and #3.

For example, in the case of heart K_p prediction in group 3, it was observed that Berezhkovskiy [12] under-predicted and Schmitt [6] over-predicted the K_ps (Figure 3-10). Classification tree #3 for heart resulted in a good predictive performance with the standard deviation of log(pred/obs) being close to zero (Figure 3-10). As well, Classification tree #3 had a lower tissue specific RMSE of 0.36 compared to the other three algorithms (Table 3-17). This indicated that the classifier both performed well in classification with a rate of correct classification of 0.56 and improved the K_p prediction accuracy with RMSE of 0.45 (Table 3-16). This case is an example that supports the hypothesis that the use of a Classification tree improves K_p prediction accuracy.

Comparison of Classification tree #1, #2 and #3

When experimentally determined muscle K_p along with physicochemical parameters (e.g. LogP, pKa, and fup) are available, 6 K_p prediction algorithms can be used and these were the algorithms used in Classification tree #1. It was observed that the use of Classification tree #1 improved the K_p prediction accuracy over any one of the 6 prediction algorithms and resulted in a lower global RMSE and a higher percentage within K-fold deviation from the observed K_ps (Table 3-12, Figure 3-5).

Both the Yun and Edginton algorithm [19] and Classification tree #2 had a high K_p prediction accuracy with a high percentage within K-fold deviation from the observed K_ps. Notably, both the Jansson *et al*.

[5] and Yun and Edginton [19] models that used Vss had high accuracy and precision in K_p prediction. This further implies that the availability of the *in vivo* parameter Vss and the use of these correlation models improve K_p prediction accuracy over TCB algorithms. For the most part, the high prediction accuracy with low global RMSE may be due to their good predictive performance in bases. It was observed that about 27% and 16% of K_ps of basic compounds were best predicted by Yun and Edginton [19], and Jansson *et al*. [5] (Figure 3-4) respectively. This predictability might have led to the small global RMSE.

TCB models [6,8,9,12] only require a minimal number of input parameters such as *ex vivo* fup and physicochemical parameters. Classification tree #3 identified the best predicting model based on the basic parameters (pKa, fup, LogP) and improved the K_p prediction accuracy over any one TCB prediction algorithm alone. It is expected that Classification tree #3 will be the most applicable in early drug discovery when compared to Classification tree #1 and #2. This is because the use of the Classification tree #1 and #2 is limited by the availability of an *in vivo* parameter (i.e. muscle K_p or Vss). As discussed in the section 2.5 Discussion, correlation-based models are dependent on the dataset that is used in their derivation. The correlation model may perform better if the chemical properties of the new compound are similar to the chemical properties that were used for the development of the regression equations. This is only true if the chemical properties are only the determinants for tissue distribution of the compound. In the case where the chemical properties of the new drug are not similar to the chemical properties that were used for the development of the regression equations, a TCB model may perform better than a correlation model. This is because a TCB model is not empirical but mechanistic. Therefore, the performance of K_p prediction algorithms should be evaluated using an external dataset that was not used for the development of the correlation model because the prediction performance of a regression-based algorithm could be artificial depending on the dataset. Recently, researchers compared

the predictive performance of K_p algorithms using Vss as an outcome. Using an independent dataset [34] it was found that a correlation model (i.e. Jansson *et al*. [5]) had better K_p prediction performance than a TCB model (i.e. Rodgers *et al*. [8,9]). However, the TCB models do have an advantage in that they are applicable for any species if the tissue-specific physiological parameters are available. For regression based algorithms that were built using rat *in vivo* or *ex vivo* data, the ratio of rat to the species of interest fup have been used for inter-species scaling [7].

For the most part, Classification trees had better prediction performance in most tissues (Figure 3-6, Figure 3-8, Figure 3-10) with little bias towards over- or under-prediction (Figure 3-6). According to the plots of the percentage of predicted K_ps within 1.25 and 1.5 fold deviations from the observed K_ps, Classification trees #1, #2 and #3 had higher percentages when compared to other algorithms in each group. Based on these results, it can be concluded that Classifications trees offer advantages over using any single algorithm to predict all tissue-specific K_ps for a compound.

Limitations of current K_p *prediction algorithms*

The accuracy of the TCB method depends on how well the factors describing the underlying process in tissue distribution (e.g. compound binding affinity to cell constituents) are formulated. Unreasonable formulation in the structure or uncertainty in physiological and/or chemical parameter values can lead to poor prediction in K_p. An underlying mechanism of a K_p prediction algorithm may not be true for a compound in certain physicochemical space. For example, a different approach was needed to overcome the poor K_p prediction accuracy for highly lipophilic compounds. It is known that the high lipophilicity of a compound is associated with a large tissue distribution (i.e. large K_p, large Vss). Rodgers *et al*. [77] demonstrated that Vss increases exponentially when LogP increases above a LogP of 6. In terms of the currently available algorithms (e.g. Jansson et al., Rodgers et al., Yun and Edginton), all equations are

designed such that an increase in lipophilicity leads to the increase in K_p values. Above a certain LogP value, however, this relationship between distributional parameters and LogP may not hold true as K_p and/or Vss may reach a plateau [15,78]. Therefore, in Poulin and Haddad's simplified model [79] for highly lipophilic compounds (logP > 6), regardless of a compound's acid-base-neutral properties, compound partitioning into neutral lipids is prevalent [79] and the plateau concept holds true. In the present study, the range of LogP values was -3 to 6. This means that all of the algorithms included in the Classification trees are not appropriate to use with compounds where LogP is greater than 6. Therefore, user caution is recommended for K_p prediction of highly lipophilic compounds (LogP > 6). As drug compounds tend to have LogP values less than 6, this is not expected to affect the accuracy of small drug molecule K_p prediction. For environmental contaminants however, LogP values often exceed 6 and the use of certain algorithms will over-predict K_ps.

In the presence of transport carriers, there would be a discrepancy between true K_p and the estimated K_p under the assumption of no carrier mediated tissue partitioning. The empirical model for estimating K_ps is highly dependent on the development dataset. If a dataset is comprised of numerous compounds for which tissue distribution is affected by active transport, those observations in the dataset can be influential in determining the coefficient of an equation which can lead to the poor K_p prediction of a new observation. The relationship between *in vivo* parameters, chemical properties of a compound and tissue K_ps is not currently robust enough to describe the tissue partitioning in the presence of carrier-mediated distribution. Thus, user discretion is recommended in the use of K_p prediction algorithms for compounds that are significantly affected by elimination and active transport. Despite this limitation, the predictive performance of the proposed algorithm was evaluated. It was found that the proposed algorithm had higher tissue-specific prediction accuracy than previously published K_p prediction algorithms in most tissues.

One of the advantages of K_p prediction algorithms is to provide an estimation of K_ps based on physiological and physicochemical parameters without experimental determination in animals. A K_p prediction algorithm is a simplified model (i.e. assumption of passive diffusion of compounds) and may overlook important biological processes (such as elimination or carrier mediated distribution). However, in the process of building a PBPK model, this passive diffusion K_p is the desired input parameter. The effect of extensive metabolism in an eliminating organ or the effect of transporters in tissue distribution is taken into account, not through a K_p, but through the incorporation of the enzyme or transporters.

3.6 Conclusion

The Classification tree based K_p prediction requires readily available parameters such as LogP, pKa, fup, and *in vivo* parameters (i.e. a muscle K_p or Vss). Classification trees have the advantage of using the best predicting algorithm for a compound within a specific tissue. Each algorithm has its unique theory in the K_p prediction and different underlying processes are previously described (Chapter 1 Introduction). For example, some algorithms put more emphasis on the fact that electrostatic binding of basic compounds to phosphophatidylserine mainly drives tissue partitioning. Other algorithms focus on the relationship between muscle K_p and lean tissue K_ps, and predictive regression equations were derived using this relationship. Based on readily available compound-specific parameters, the Classification tree classified and identified which algorithm best described the tissue partitioning for a compound. As a result, the Classification tree based K_p prediction improved accuracy over using any one K_p prediction algorithm.

Chapter 4

Conclusions and future work

Tissue-to-plasma partition coefficients (K_p) that characterize the tissue distribution of a compound are important input parameters in PBPK models. This study proposed two different approaches for K_p prediction. Predictive regression equations that use readily available parameters were developed. This approach is computationally simple, but the use is limited to the availability of the *in vivo* parameter of Vss. It was found that the developed regression equations had greater prediction accuracy in comparison to published K_p prediction algorithms.

In terms of the Classification tree based K_p prediction method, the use of previously published algorithms and the identification of the most accurate algorithms resulted in a competitive K_p prediction over any one algorithm alone. This was particularly evident with Classification tree #3 that identified the best tissue composition model and greatly improved *a priori* K_p prediction. In the absence of *in vivo* data (i.e. muscle K_p and Vss), Classification tree #3 had better predictive performance when compared to using a single TCB model.

One of the limitations of the Classification tree based K_p prediction is that it is mathematically complicated. In order to overcome this problem, the Classification trees will be available as a web based program for public consumption as a future work. This will feature the Classification tree calculator that will define the best predicting algorithm as well as a K_p calculator for calculating K_p from the best predicting algorithm. This program will be used as a tool for K_p prediction and requires only a minimal number of input parameters (i.e. LogP, pKa, fup, Vss and/or muscle K_p).

In conclusion, this study proposed an improved K_p correlation algorithm and a novel Classification tree that led to a more accurate K_p prediction. Classification tree based K_p prediction overcomes the

limitations of any one algorithm by harnessing the best components of each algorithm. The predictive performances of the two methods were demonstrated to be superior to previously published K_p algorithms. An accurate prediction of target site concentrations is of great importance as this concentration drives pharmacological response. Increased prediction accuracy of K_ps will lead to the appropriate parameterization of PBPK models and will enhance the predictability of a compounds' pharmacokinetics.

Appendix A

Appendix 1. Development set A of moderate to strong bases to construct a predictive regression equation.

Drug	LogP	pKa	Drug Class[a]	fup	Vss (L/Kg)	Adipose	Bone	Brain	Gut	Heart	Kidney	Liver	Lung	Muscle	Skin	Spleen
Acebutolol-R [80]	1.79	9.7	B	0.79	9.33	1.10	0.06	0.48	22.43	5.71	23.58	31.48	10.31	4.97	3.01	
Acebutolol-S [80]	1.79	9.7	B	0.73	8.90	0.79	0.04	0.36	91.25	4.30	32.70	24.89	6.14	4.45	2.47	
Betaxolol-R [80]	2.59	9.4	B	0.53	20.99	2.95	13.20	12.93	40.23	23.59	58.30	130.91	203.52	13.78	6.52	
Betaxolol-S [80]	2.59	9.4	B	0.54	19.75	2.86	12.85	13.01	37.80	21.52	54.54	108.00	182.52	13.55	6.05	
Bisoprolol-R [80]	1.87	9.4	B	0.85	6.92	1.03	4.88	1.64	26.52	6.49	24.91	22.78	41.82	5.40	2.18	
Bisoprolol-S [80]	1.87	9.4	B	0.85	6.72	1.02	4.43	1.79	25.67	6.69	24.82	22.95	41.99	5.23	2.21	
Caffeine [81]	0.17	10.4	B	0.97	0.71	0.23	0.89	0.60		0.56	0.93					
Carvedilol-R [80]	4.19	8.1	B	0.02	1.79	0.80				1.94	1.92	4.52	34.00	0.81		
Chlorpromazine [2]	5.42	9.7	B	0.11	29			11.50								
Cocaine [2]	2.30	8.6	B	0.63	2.80	5.16		7.02	6.94		13.18			3.02		
Cotinine [2,29]	-0.25	8.1	B	0.97	0.43	0.08		0.42	0.64	0.51	0.99	0.64	0.63	0.67		
Haloperidol [2]	4.30	8.7	B	0.23	10		27.20	13.37	10.80	14.30			53.50	29.00	6.20	
Inaperisone [82]	3.50	9.0	B	0.24	6.35	16.00		12.00		7.40	58.00	34.00	33.00	4.10	6.30	
Lidocaine [2]	2.44	8.0	B	0.38	2.62			3.24	3.12	2.73	17.21	11.51	3.80	1.68	2.58	4.79
Metoprolol-R [80]	2.01	9.7	B	0.80	7.87	1.04	5.18	6.48	12.96	6.89	26.56	40.04	25.56	5.66	3.19	
Metoprolol-S [80]	2.01	9.7	B	0.81	7.74	0.98	5.33	6.97	11.22	6.25	26.89	44.59	26.57	5.57	2.92	
Morphine [83]	0.82	8.3	B	0.72	5.18						9.50	1.20		2.50		
Nicotine [84]	1.17	7.8	B	0.84	1.53	0.32		2.02	1.60	1.12	18.14	4.95	1.24	1.23	1.10	

Oxprenolol-R [80]	2.18	9.5	B	0.24	2.80	0.58	1.87	1.29	12.71	3.62	14.17	8.59	15.86	3.08	1.37	
Oxprenolol-S [80]	2.18	9.5	B	0.36	3.74	0.69	2.33	2.48	11.18	4.37	17.62	12.33	21.24	3.89	1.70	
Pentazocine [2]	3.31	8.5	B	0.46	7.66	2.50	5.40	4.30	4.70	5.40	20.00	2.30	27.00	5.90	4.70	
Pethidine [80]	2.45	8.6	B	0.15	13.20	4.17			16.60			262.28	24.24	5.20		
Pindolol-R [80]	1.75	9.0	B	0.51	4.32	0.88	2.71	5.10	26.01	13.87	47.40	14.36	33.58	8.08	2.86	
Pindolol-S [80]	1.75	9.0	B	0.76	8.59	0.62	2.29	5.17	18.32	9.27	29.79	7.24	30.32	7.28	2.74	
Procainamide [2]	0.88	9.2	B	0.92	1.77	0.13		2.47		2.48	6.38	3.19		4.38		
Propranolol [2]	3.22	9.4	B	0.08	13.04			14.00	6.60	7.10	15.30	11.60	16.46	4.30		14.20
Propranolol-R [80]	3.48	9.5	B	0.02	1.88	0.65	1.39	6.51	6.27	3.86	6.19	5.56	24.24	1.89	1.09	
Propranolol-S [80]	3.48	9.5	B	0.13	10.13	2.41	6.73	35.69	23.11	15.75	35.31	29.34	131.70	9.40	5.21	
Pyridostigmine [2]	-3.73	10	B	0.50	0.35					1.10	15.20	2.10		0.52		
Theophyllin [2]	0.26	8.7	B	0.60	0.95			0.36					0.71	0.60		
Verapamil [2,85]	3.79	8.5	B	0.05	4.40					6.00	12.50		50.00	3.50		
Quinidine [86]	3.40	9.3	B	0.33	8.94			1.16	14.42	8.92	19.51	20.79	44.03	3.82		23.99
Timolol-S [80]	1.87	9.2,8,8	BZ	0.63	5.20	0.64	1.00	1.06	20.16	5.36	13.32	7.87	26.96	4.15	1.58	
Enoxacin [87]	0.10	8.7,6.1	BZ	0.66	1.57		1.44			1.07	4.61	3.21	1.14	1.45	1.36	1.63
Ofloxacin [87]	-0.40	8.2,6.1	BZ	0.77	1.50	0.19	1.42	0.24		1.78	6.39	2.04	1.36	1.72	1.19	1.93
Tetracycline [88]	0.03	9.7,7.7,3.3	BZ	0.50	2.20	1.10	8.11		3.75		4.05	4.70		1.62		
Pefloxacin [87]	0.42	7.6,6.3	BZ	0.77	2.75			0.16		2.36	4.13	5.34	1.94	2.41		3.42
JNJ1/Domperidone [89]	3.96	7.9	B	0.09	7.40	3.21		0.12		3.87	22.50	13.80	10.90	3.45	4.35	
JNJ13/Prucalopride [89]	2.26	8.5	B	0.71	4.90			0.43		4.30	17.60	8.77	10.60	4.57		
JNJ14/Sabeluzole [89]	4.63	7.8	B	0.02	5.85	8.41	1.83	5.37		2.45	10.40	37.70	29.20	0.83	2.95	5.48
JNJ15/Lubeluzole [89]	4.88	7.6	B	0.01	4.24			4.13			9.90	27.70	18.10	2.04		
JNJ18/Laniquidar [89]	5.50	7.9	B	0.00	8.95			2.86		5.82	12.00	16.80	38.70	7.07		
JNJ2/Nebivolol [89]	4.03	8.4	B	0.02	5.20			3.73		4.71	10.60	14.10	99.70	2.95		

JNJ28/Sufentanil [89]	4.02	8.1	B	0.07	4.32	7.72		2.08		1.80	1.17	0.37	6.18	1.71		2.80
JNJ29 [89]	4.18	8.9	B	0.06	12.90			11.60		13.70	25.30	63.80	122.00	8.00		
JNJ3/Galantamine [89]	1.09	8.2	B	0.76	5.18			1.51			13.90	2.53		2.14		
JNJ30 [89]	4.90	7.7	B	0.02	7.11			1.26		2.61	18.10	12.00	47.70	7.73		
JNJ33 [89]	2.08	8.3	B	0.63	3.00			0.24		3.00	13.80	8.90	7.80	2.60		
JNJ37 [89]	4.60	9.1	B	0.04	32.70			34.00		36.00	44.00	212.00	297.00	14.00		
JNJ6/Loperamide [89]	5.13	8.9	B	0.02	4.42						9.30	5.00	35.90			
JNJ7 [89]	2.47	7.8	B	0.53	3.28					4.44	18.10	31.00	11.70	4.40		
JNJ8 [89]	1.18	9.9	B	0.82	7.08					5.15	29.70	45.90	12.20			
JNJ9/Cisapride [89]	4.22	7.9	B	0.08	4.73			1.56		1.93	7.32	17.10	10.80			
Ketanserin [90]	3.30	7.5	B	0.01	0.67	0.56	0.19	0.19		0.35	1.53	2.60	1.49	0.28	0.46	0.91
Risperidone [90]	3.04	8.2	B	0.12	1.77			0.23		0.82	0.64	12.30	3.42			
Levocabastine [90]	1.75	9.3,3.2	BZ	0.47	1.36	0.84	0.52	0.59		1.19	8.52	14.00	1.49	0.88	0.98	1.32
Norfloxacin [2]	-1.03	8.8,6.6	BZ	0.58	2.05								1.34	0.92		
Grepafloxacin [91]	1.17	9.08,6.08	BZ	0.59	5.42				6.06	5.19	15.01	11.46	20.23	3.54		
Sparfloxacin [92]	0.21	9.08,5.84	BZ	0.55	3.42	0.18		0.00	9.87	2.09	7.55	4.50	2.45	1.93	2.08	

[a]B:base, BZ: polyproritic compound with basic pKa ≥7.4

Appendix 2. Development set B of acids, neutrals and weak bases to construct a predictive regression equation.

Drug	LogP	pKa	Drug Class[a]	fup	Vss (L/Kg)	Adipose	Bone	Brain	Gut	Heart	Kidney	Liver	Lung	Muscle	Skin	Spleen
Penicillin [93]	1.64	2.8	A	0.15	0.24				0.97	0.10	3.71	0.25	0.16	0.06		0.10
Salicylic acid [94]	2.26	4.0	A	0.40	0.19		0.14	0.06	0.66	0.19	0.44	0.23	0.19	0.13	0.27	
Valproic acid [9]	2.75	4.6	A	0.37	0.66	0.15		0.07	0.45	0.43	1.50	1.80	0.42	0.16	0.47	
Glycyrrhizin [2]	2.80	5.3	A	0.05	0.06					0.25			0.06	0.06	0.15	0.07
Tenoxicam [92]	1.86	5.3	A	0.02	0.13	0.02	0.08	0.01	0.17	0.14	0.78	0.86	0.24	0.06	0.12	0.07
Fleroxacin [2]	0.24	6.5	A	0.75	1.30		1.20			2.55			2.00	2.00	1.20	
5-hexyl-5-ethyl barbituric acids [95]	2.79	7.7	A	0.19	0.94	6.14		1.66	1.61	1.56	2.28	3.34	1.07	1.20	2.13	0.84
5-n-Ethyl-5-ethyl barbituric acids [95]	0.68	7.8	A	0.95	0.51	0.42	0.63	0.68	0.59	0.73	1.71	1.64	0.84	0.70	0.77	0.52
5-propyl-5-ethyl barbituric acid [95]	0.77	7.8	A	0.87	0.56	0.77	1.30	0.91	0.81	1.03	2.81	1.68	1.12	0.90	1.00	0.53
5-octyl-ethyl-barbituric acid [95]	3.82	7.8	A	0.00	0.44	5.13		1.87	1.32	1.47	2.52	3.47	3.06	0.80	1.91	1.87
5-n-heptyl-5-ethyl barbituric acids [95]	3.64	7.8	A	0.07	0.56	5.55		1.13	1.34	1.33	2.05	2.23	1.20	0.90	1.46	1.25
5-n-butyl-5-ethyl barbituric acids [95]	1.70	7.8	A	0.61	0.57	1.31	0.98	1.17	1.23	1.45	3.24	2.09	1.05	0.90	1.09	0.36
5-nonyl-5-ethyl barbituric acid [95]	4.07	7.8	A	0.01	1.34	5.83		2.49	2.04	2.07	4.07	3.76	2.65	1.00	2.76	3.09
5-pentyl-5-ethyl barbituric acid [95]	2.20	8.0	A	0.50	0.74	1.63	0.49	0.91	0.82	0.91	2.27	1.72	0.65	0.70	1.11	0.33
Hexobarbital [29,96]	1.74	8.1	A	0.70	1.20	1.60			1.43	1.28	1.50	6.00	2.81	1.00	0.95	
5-n-Methyl-5-ethyl barbituric acids [95]	0.05	8.1	A	1.00	0.71	0.27	0.98	0.63	0.59	0.68	1.30	1.50	0.73	0.60	0.76	0.70
Phenytoin [4,97]	2.47	8.2	A	0.12	1.39	1.64		0.70	1.24	0.71	1.60	2.30	0.72	0.70	0.94	
Nalidixic acid [87]	1.10	5.1,3.3	Z	0.29	0.38		0.29	0.22	0.49	0.49	0.54	0.58	0.33	0.36	0.35	0.00
Ftorafur [9]	-0.27		N	0.78	0.34	0.17		0.41	0.36	0.38	0.68	0.39	0.26	0.50	0.40	0.42
2,3-Dideoxyinosine [2]	-1.24		N	0.98	0.51			0.46	0.51		6.86	0.77		0.69		0.96
Ethoxybenzamide [2]	0.80		N	0.59	0.63	0.71		0.94	0.56	0.99	1.30		0.91	0.81	1.04	0.87

Digoxin [2]	1.23		N	0.73	0.99				5.91	1.65	2.07	15.19	2.09	1.40		
Prednisolone [2]	2.02		N	0.23	1.37			0.48		0.67			0.66	0.35		
Clobazam [2]	1.84		N	0.25	3.29									2.60		
Cyclosporin [98]	2.90		N	0.08	3.62	11.57	3.18	0.79	5.23	4.05	7.99	12.20	5.52	1.35	2.92	5.45
Propofol [2]	3.79		N	0.03	9.90			8.20		4.33		13.07	4.41	1.06		
Triazolam [3]	2.40		N	0.28	2.24	6.02	0.00		11.90		8.43	3.75		6.02	5.46	
Alprazolam [3]	2.21		N	0.35	1.98	1.08	0.95	1.88	1.67	1.69	3.68	8.39	3.15	2.00	2.96	
Chlordiazepoxide [3]	2.40		N	0.15	1.45	4.31	0.00	0.75	1.97	2.61	2.70	4.85		0.77	0.48	
Midazolam [3]	3.01	5.9	WB	0.04	2.38	4.62	1.92	2.49	2.81	4.64	3.19	8.51	4.08	0.87	1.96	2.42
JNJ17 [89]	7.00	6.8	WB	0.02	6.94			0.79		4.84		11.70	20.60	2.95		
JNJ20 [89]	3.23	7.0	WB	0.08	1.58			1.34		1.45	4.47	7.44		0.67		
JNJ23 [90]	3.40	7,3.1	WB	0.08	1.58	2.53	0.69	1.39		1.34	4.03	8.64	2.48	0.67	0.92	3.35
JNJ25 [89]	4.43	7.2	WB	0.04	6.47			0.63		3.39	8.25	21.40	22.90	1.47		
JNJ21 [90]	4.17	7.2	WB	0.01	7.35			1.15		1.53	2.95	15.90	3.49	0.49		
JNJ24 [89]	4.69	7.3	WB	0.02	10.70			4.55		7.41		20.90		4.50		
Ridogrel [90]	3.54	4.9,3.8	Z	0.05	0.78			0.18		0.39	0.25	1.39	0.37	0.11		

[a]A: acid, N: neutral, WB: weak base, Z: zwitterion

Appendix 3. Test set A for moderate to strong bases to evaluate prediction accuracy.

Drug	LogP	pKa	Drug Class[a]	fup	Vss (L/Kg)	Adipose	Bone	Brain	Gut	Heart	Kidney	Liver	Lung	Muscle	Skin	Spleen
Biperiden [2]	4.25	8.8	B	0.17	14.00	67.64	2.28	7.95	12.92	8.04	12.13		86.31	3.69	4.70	
Carvedilol-S [80]	4.19	8.1	B	0.04	3.36	1.90				7.42	7.00	11.77	75.60	1.60		

Fentanyl [53,99]	3.97	8.7	B	0.16	4.58	26.70		3.53	8.36	4.50	12.09	3.80	13.50	3.09	2.09	27.60
JNJ4/Lorcainide [89]	4.16	9.4	B	0.26	3.92	5.27		1.52		2.90	5.67	0.57	19.40	2.82		
Imipramine [53]	4.62	9.5	B	0.24	18.69	7.35		22.99	26.66	21.91	54.19	121.28	141.22	9.91	1.68	57.36
Phencyclidine [2]	4.96	9.4	B	0.47	12.55	61.57		2.57		2.19	11.80	8.04	40.98	1.51		
Lomefloxacin [87]	-0.30	9.3	B	0.72	1.30	0.27	1.58	0.22	1.63	1.37	4.84	2.30	1.24	1.61	0.94	1.73
Pipemidic acid [87]	-2.15	7.5,4.9	Z	0.82	2.31	0.34	2.02	0.13		0.89	7.41	4.61	1.03	1.05		1.35
Disopyramide [85]	2.58	9.4	B	0.24	0.90			0.94		2.03				2.30		
FTY-720 [100]	4.06	8.7	B	0.00	13.70			49.20		17.40	35.80	47.00	68.20	10.50		62.10

[a]B: base, Z: zwitterion

Appendix 4. Test set B for acids, neutrals and weak bases to evaluate prediction accuracy.

Drug	LogP	pKa	Drug Class[a]	Fup	Vss (L/Kg)	Adipose	Bone	Brain	Gut	Heart	Kidney	Liver	Lung	Muscle	Skin	Spleen
Thiopental [4]	2.85	7.5	A	0.18	0.19	8.00		0.70	1.32	1.40	3.09	2.29	1.54	0.88	1.18	0.53
Tolbutamide [101]	2.34	5.5	A	0.24	0.20	0.13		0.10	0.12	0.27	0.22	0.30	0.25	0.13	0.22	0.19
Cefazolin [93]	0.28	2.3	A	0.15	0.40		0.11		0.17	0.10	2.77	0.77	0.19	0.09	0.30	
Ceftazidime [102]	-0.50	3.92,2.5,1.9	Z	0.10	0.24	0.16			0.41	0.22	4.80	0.25	0.44	0.19	0.39	
Bromperidol [2]	4.03	8.0	N	0.50	10.10			24.00								
Pentobarbital [2]	2.10	8.1	A	0.66	1.30	1.30								0.80		
Flunitrazepam [3]	2.34	1.8	N	0.25	4.54	73.50	4.36	1.46	2.74	1.66	0.40	3.69	4.78	1.03		
Mazapertine [90]	5.05	7.0	WB	0.03	3.15	8.01		0.62		1.52	7.36	20.50	2.31	1.49	1.12	1.55
Alfentanil [99]	2.20	6.5	WB	0.16	0.71	1.89		0.13	1.18	0.55	0.82	1.00	0.78	0.31	0.18	0.73
Diazepam [3]	2.87	3.4	WB	0.13	5.12	12.20	5.45	2.13	7.06	5.56	4.15	13.44	5.89	2.77	4.23	

[a]A: acid, N: neutral, WB: weak base, Z: zwitterion

Appendix 5. Sample R codes Random forest, bagging and Rpart

```
rm(list = ls(all = TRUE))
library(MASS)
library(RODBC)
channel <- odbcConnectExcel()
mydata <- sqlFetch(channel, "Heart")
odbcClose(channel)
tr<-mydata
logp<-tr$LogP
fup<-tr$fup
doi<-tr$DOI7#1
group<-as.factor(tr$Code)
muscle<-tr$Muscle
trdata<-data.frame(logp,fup,doi,muscle)
library(randomForest)
rf<-randomForest(group~.,data=trdata,na.action=na.omit)
library(ipred)
bag<-bagging(group~.,data=trdata)
library(rpart)
rpart<-rpart(group~.,data=trdata,method="class")
plot(rpart, compress=T,uniform=T,margin=0.1)
text(rpart, use.n=T,col='blue')
printcp(rpart)
plotcp(rpart)
pfit<- prune(rpart, cp=   rpart$cptable[which.min(rpart$cptable[,"xerror"]),"CP"])
# pruning the tree with optimal Cp
ts<-read.csv("DT1-ts-Oct17-final.csv")
logp<-ts$LogP
fup<-ts$fup
doi<-ts$DOI7#1
muscle<-ts$Muscle
tsdata<-data.frame(logp,fup,doi,muscle)
table(predict(rf,tsdata,na.action=na.omit))
rfresult<-data.frame(predict(rf,tsdata,na.action=na.omit))
rfresult
table(predict(bag,tsdata))
bagresult<-data.frame(predict(bag,tsdata))
bagresult
table(predict(rpart,tsdata),ts$group)
rpartresult<-data.frame(predict(pfit,tsdata))
rpartresult
printcp(rpart)
plotcp(rpart)
```

Appendix 6. Sample R codes for generation of final Classification trees by random forest analysis with the total dataset

```
rm(list = ls(all = TRUE))
library(stats) # calling stats library
tr<-read.csv("DEC15-DT2-code-Final.csv",sep=",") #reading the dataset
logp<-tr$LogP # reading dataset for variables
fup<-tr$fup
doi<-tr$DOI7 #Degree of ionization at pH 7
vss<-tr$Vss
Class<-tr$Class
group<-as.factor(tr$Code_Spleen) # making the membership as a factor variable
trdata<-data.frame(logp,fup,doi,vss,Class,group) # making data frame of the training set
trdata<-na.omit(trdata)
group<-as.factor(trdata[,6])
trdata
library(randomForest) # calling randomForest library
rf<-randomForest(group~.,data=trdata[,-6]) #making random forest
rf # show result of random forest
result <- rfcv(trdata[,-6], group,cv.fold=20) # finding optimal m_try by random forest cross validation
result
cv<-data.frame(group,result$predicted$`5`) # compare the true classification and the classification by random forest
cv #show result
```

Appendix 7. Dataset for random forest analysis; summary of compound specific physicochemical parameters

Compound	LogP	pKa	Drug Class[a]	fup	Vss_rat(L/Kg)	Groups		
2,4-Dichlorophenoxyacetic acid [2]	2.43	2.98	A	0.05				3
Glycyrrhetinic acid [2]	5.50	4.71	A	0.05		1		3
5-hexyl-5-ethyl barbituric acid [95]	2.79	7.74	A	0.19	0.94	1	2	3
5-n-butyl-5-ethyl barbituric acids [95]	1.70	7.81	A	0.61	0.68	1	2	3
5-n-Ethyl-5-ethyl barbituric acids [95]	0.68	7.75	A	0.95	0.51	1	2	3
5-n-heptyl-5-ethyl barbituric acids [95]	3.64	7.78	A	0.07	1.10	1	2	3
5-n-Methyl-5-ethyl barbituric acids [95]	0.05	8.11	A	0.99	0.57	1	2	3
5-nonyl-5-ethyl barbituric acid [95]	4.07	7.82	A	0.01	1.90	1	2	3
5-octyl-ethyl-barbituric acid[95]	3.82	7.78	A	0.00	1.40	1	2	3
5-pentyl-5-ethyl barbituric acid[95]	2.20	8.00	A	0.50	0.74	1	2	3
5-propyl-5-ethyl barbituric acid[95]	0.87	7.77	A	0.87	0.56	1	2	3
Cefazolin[93]	-0.58	2.28	A	0.15	0.40	1	2	3
Dicloxacillin[93]	2.91	2.88	A	0.03		1		3
Etodolac-R[103]	3.60	4.70	A	0.00				3
Etodolac-S[103]	3.60	4.70	A	0.02				3
Fleroxacin[2]	0.24	6.50	A	0.75	1.30	1	2	3
Glycyrrhizin[2]	2.80	5.30	A	0.05	0.18	1	2	3
Hexobarbital[2]	1.74	8.10	A	0.70	1.20	1	2	3
Penicillin [93]	1.64	2.80	A	0.15	0.24	1	2	3
Phenobarbital [2]	1.47	7.35	A	0.78	1.02	1	2	3
Phenytoin[4,97]	2.47	8.23	A	0.12	1.39	1	2	3
p-Phenylbenzoic acid[104]	2.81	4.20	A	0.03		1		3
Salicylic acid [94]	2.26	3.00	A	0.40	0.19	1	2	3
Tenoxicam [9]	1.86	5.30	A	0.02	0.13	1	2	3
Thiopental[2]	2.85	7.50	A	0.18	0.19	1	2	3
Tolbutamide [2]	2.34	5.50	A	0.24	0.20	1	2	3
Valproic acid [2]	2.75	4.60	A	0.37	0.66	1	2	3
Caffeine[5]	1.29	10.40	B	0.97	0.71		2	3
Chlorpromazine [2]	5.42	9.70, 6.40	B	0.11	29.00		2	3
Cocaine[5]	2.30	8.61	B	0.63	2.80	1	2	3
Disopyramide R- [85]	2.71	9.92	B	0.24		1		3
Disopyramide S- [85]	2.71	9.92	B	0.24		1		3
Flecainide R- [85]	4.65	9.80	B	0.52		1		3
Flecainide S- [85]	4.65	9.80	B	0.52		1		3
Flurazepam [2]	3.80	9.79	B	0.50		1		3
N-Acetylprocainamide [2]	1.50	9.09	B	0.92				3
Pethidine [105,106]	2.45	8.59	B	0.15	13.20	1	2	3
Phencyclidine [8]	4.96	9.40	B	0.47	12.55	1	2	3

Trihexyphenidyl [2]	4.30	8.70	B	0.37		1		3
Verapamil R- [85]	3.79	8.92	B	0.10		1		3
Verapamil S- [85]	4.92	8.92	B	0.10		1		3
Domperidone [89]	3.96	7.89	B	0.09	7.40	1	2	3
Nebivolol [89]	4.03	8.40	B	0.02	5.20	1	2	3
Galantamine [89]	1.09	8.20	B	0.76	5.18	1	2	3
Lorcainide [89]	4.16	9.44	B	0.26	4.59	1	2	3
Fentanyl [89]	3.94	8.40	B	0.17	3.65	1	2	3
Loperamide [89]	5.13	8.86	B	0.02	4.42		2	3
Cisapride [89]	4.22	7.90	B	0.08	4.73		2	3
Ritanserin [89]	5.20	8.20	B	0.02	8.00	1	2	3
Prucalopride [89]	2.26	8.50	B	0.71	4.90	1	2	3
Sabeluzole [89]	4.63	7.80	B	0.02	5.85	1	2	3
Lubeluzole [89]	4.88	7.60	B	0.01	4.24	1	2	3
Laniquidar [89]	5.50	7.90	B	0.00	8.95	1	2	3
Acebutolol-R [80]	1.79	9.70	B	0.79	9.33	1	2	3
Acebutolol-S [80]	1.79	9.70	B	0.73	8.90	1	2	3
Betaxolol-R [80]	2.59	9.40	B	0.53	20.99	1	2	3
Betaxolol-S [80]	2.59	9.40	B	0.54	19.75	1	2	3
Biperiden [2]	4.25	8.80	B	0.17	14.00	1	2	3
Bisoprolol-R [80]	1.87	9.40	B	0.85	6.92	1	2	3
Bisoprolol-S [80]	1.87	9.40	B	0.85	6.72	1	2	3
Carvedilol-R [80]	4.19	8.10	B	0.02	1.79	1	2	3
Carvedilol-S [80]	4.19	8.10	B	0.04	3.36	1	2	3
Clozapine [2]	3.23	7.50	B	0.50				3
Cotinine [2]	-0.25	8.10	B	0.97	0.43	1	2	3
Diazepam [3]	2.87	3.40	B	0.15	5.12	1	2	3
Haloperidol [2]	4.30	8.70	B	0.23	10.00	1	2	3
Imipramine [53]	4.62	9.50	B	0.24	18.69	1	2	3
Inaperisone [82]	3.50	8.97	B	0.24	6.35	1	2	3
Lidocaine [2]	2.44	8.00	B	0.38	2.62	1	2	3
Metoprolol-R [80]	2.01	9.70	B	0.80	7.87	1	2	3
Metoprolol-S [80]	2.01	9.70	B	0.81	7.74	1	2	3
Morphine [83,107]	0.82	8.28	B	0.72	5.18	1	2	3
Nicotine [2]	1.17	7.80, 3.00	B	0.84	1.53	1	2	3
Oxprenolol-R [80]	2.18	9.50	B	0.24	2.80	1	2	3
Oxprenolol-S [80]	2.18	9.50	B	0.36	3.74	1	2	3
Pentazocine [2]	3.31	8.50	B	0.46	7.66	1	2	3
Pindolol-R [80]	1.75	9.05	B	0.51	4.32	1	2	3
Pindolol-S [80]	1.75	9.05	B	0.76	8.59	1	2	3
Procainamide [2]	0.88	9.20	B	0.92	1.77	1	2	3

Promazine [2]	4.55	9.10	B	0.05				3
Propranolol [2]	3.22	9.41	B	0.08	13.04	1	2	3
Propranolol-R [80]	3.48	9.50	B	0.02	1.88	1	2	3
Propranolol-S [80]	3.48	9.50	B	0.13	10.13	1	2	3
Pyridostigmine [2]	-3.73	10.00	B	0.50	0.35	1	2	3
Quinidine [86]	3.01	10.00, 5.40	B	0.33	8.94	1	2	3
Theophyllin [2]	-0.02	8.81	B	0.90	0.50	1	2	3
Thioridazine [2]	5.90	9.50	B	0.01				3
Timolol-S [80]	1.87	9.20, 8.80	B	0.63	5.20	1	2	3
Verapamil [6]	3.79	8.50	B	0.05	4.40	1	2	3
Bromperidol [2]	4.03		N	0.50	10.10		2	3
Fluphenazine [2]	4.20		N	0.50				3
Ftorafur [9]	-0.27		N	0.78	0.34	1	2	3
Medazepam [2]	3.89		N	0.50		1		3
Neostigmine [2]	-1.65		N	0.50		1		3
N-Methylpentobarbital [2]	2.69		N	0.50		1		3
Propofol [2]	3.79		N	0.03	9.90		2	3
2,3-Dideoxyinosine [2]	-1.24		N	0.98	0.51	1	2	3
Clobazam [2]	2.86		N	0.25	3.29	1	2	3
Cyclosporin [98]	2.90		N	0.12	3.62	1	2	3
Digoxin [2]	1.23		N	0.73	0.99	1	2	3
Ethoxybenzamide [2]	0.80		N	0.59	0.63	1	2	3
Chlordiazepoxide [9]	2.40	4.70	WB	0.15	1.45	1	2	3
Prazepam [2]	3.73	3.44	WB	0.50		1		3
Triazolam [3]	2.40	2.00	WB	0.28	2.24	1	2	3
Alfentanil [2]	2.20	6.50	WB	0.16	0.71	1	2	3
Alprazolam [3]	2.21	2.40	WB	0.35	1.98	1	2	3
Flunitrazepam [9]	2.34	1.80	WB	0.25	3.81	1	2	3
Midazolam [3]	3.01	5.87	WB	0.07	2.38	1	2	3
Sparfloxacin [92]	0.21	5.84, 9.08	Z	0.55	3.42	1	2	3
Ceftazidime [102]	-1.71	2.50,3.80, 1.90	Z	0.90	0.24	1	2	3
Nalidixic acid [87]	1.10	5.10,3.30	Z	0.29	0.38	1	2	3
Enoxacin [87]	0.10	6.10,8.70	Z	0.66	1.57	1	2	3
Lomefloxacin [87]	-0.30	5.80,9.30	Z	0.72	1.30	1	2	3
Ofloxacin [87]	-0.40	6.10,8.20	Z	0.92	1.50	1	2	3
Grepafloxacin [91]	1.17	6.08,9.08	Z	0.59	5.42	1	2	3
Norfloxacin [2]	-1.03	6.60,8.80	Z	0.58	2.05	1	2	3
Pefloxacin [2]	0.42	6.30,7.60	Z	0.77	2.75	1	2	3
Pipemidic acid [2]	-2.15	7.00,4.90,3.50	Z	0.82	2.31	1	2	3
Tetracycline [2]	-1.30	7.70,9.70,3.30	Z	0.50	2.20	1	2	3

[a]A: acid, B: base, WB: weak base with basic pKa ≤ 7.4, Z: zwitterion

Appendix 8. Dataset for random forest analysis; summary of experimentally determined K_ps

Compound	Adipose	Bone	Brain	Gut	Heart	Kidney	Liver	Lung	Muscle	Skin	Spleen
2,4-Dichlorophenoxyacetic acid [2]			1.42								
Glycyrrhetinic acid [2]			0.04		0.12			0.22	0.1	0.16	0.07
5-hexyl-5-ethyl barbituric acid [95]	6.14		1.66	1.61	1.56	2.28	3.34	1.07	1.17	2.13	0.84
5-n-butyl-5-ethyl barbituric acids [95]	1.31	0.98	1.17	1.23	1.45	3.24	2.09	1.05	0.9	1.09	0.36
5-n-Ethyl-5-ethyl barbituric acids [95]	0.42	0.63	0.68	0.59	0.73	1.71	1.64	0.84	0.66	0.77	0.52
5-n-heptyl-5-ethyl barbituric acids [95]	5.55		1.13	1.34	1.33	2.05	2.23	1.2	0.93	1.46	1.25
5-n-Methyl-5-ethyl barbituric acids [95]	0.26	0.98	0.63	0.59	0.68	1.3	1.5	0.73	0.6	0.76	0.7
5-nonyl-5-ethyl barbituric acid [95]	5.83		2.49	2.04	2.07	4.07	3.76	2.65	0.99	2.76	3.09
5-octyl-ethyl-barbituric acid[95]	5.13		1.87	1.32	1.47	2.52	3.47	3.06	0.81	1.91	1.87
5-pentyl-5-ethyl barbituric acid[95]	1.63	0.49	0.91	0.82	0.91	2.27	1.71	0.65	0.72	1.11	0.33
5-propyl-5-ethyl barbituric acid[95]	0.77	1.3	0.91	0.81	1.03	2.81	1.68	1.12	0.89	1	0.53
Cefazolin[93]		0.11		0.17	0.1	2.77	0.77	0.19	0.09	0.3	
Dicloxacillin[93]				1.4	0.07	1.3	0.43	0.12	0.05		0.09
Etodolac-R[103]	0.07		0.03		0.18	0.12	0.12				
Etodolac-S[103]	0.17		0.05		0.45	0.39	0.43				
Fleroxacin[2]		1.2			2.55			2	2	1.2	
Glycyrrhizin[2]					0.25			0.06	0.06	0.15	0.07
Hexobarbital[2]	1.6			1.43	1.28	1.5	6	2.81	0.99	0.95	
Penicillin [93]				0.97	0.1	3.71	0.25	0.16	0.06		0.1
Phenobarbital [2]	0.31		0.59	1.75	1.47	0.73	1.8	1.18	1.41	1.38	
Phenytoin[4,97]	1.64		0.7	1.24	0.71	1.6	2.3	0.72	0.7	0.94	
p-Phenylbenzoic acid[104]	0.06		0.06	0.15	0.23	0.3	0.35	0.28	0.08	0.15	0.1
Salicylic acid [94]		0.14	0.06	0.66	0.19	0.44	0.23	0.19	0.13	0.27	
Tenoxicam [9]	0.02	0.08	0.01	0.17	0.14	0.78	0.86	0.24	0.06	0.12	0.07
Thiopental[2]	15.5		0.7	1.32	2.59	3.09	2.29	2.96	2.05	1.75	0.53
Tolbutamide [2]	0.13		0.1	0.12	0.27	0.22	0.3	0.25	0.13	0.22	0.19
Valproic acid [2]	0.15		0.07	0.45	0.43	1.5	1.8	0.42	0.16	0.47	
Caffeine[5]	0.23	0.89	0.6		0.56	0.93					
Chlorpromazine [2]			11.5								
Cocaine[5]	5.16		7.02	6.94		13.18			3.02		
Disopyramide R- [85]			0.94		2.03			7.9	2.3		
Disopyramide S- [85]			0.54		2.06			7	2.13		
Flecainide R- [85]			1.5		6.75	12.8		111	7.2		
Flecainide S- [85]			1.5		6.25	16.5		76.5	6.9		
Flurazepam [2]									4.9		
N-Acetylprocainamide [2]					2.17						

Pethidine [105,106]	4.17			16.6			262.28	24.24	5.2		
Phencyclidine [8]	61.57		2.57		2.19	11.8	8.04	40.98	1.51		
Trihexyphenidyl [2]	76	7.9	21	22	23			74	13	8.1	
Verapamil R- [85]					9.4	20.9		92.2	3.8		
Verapamil S- [85]					5.6	9.4		40	0.75		
Domperidone [89]	3.21		0.12		3.87	22.5	13.8	10.9	3.45	4.35	
Nebivolol [89]			3.73		4.71	10.6	14.1	99.7	2.95		
Galantamine [89]	0.48	4.79	1.51		2.28	14.5	2.53	4.42	2.14	1.14	2.92
Lorcainide [89]	5.27		1.52		2.91	5.68	0.57	19.4	6.5		10.3
Fentanyl [89]			3.56		4.54	12.2	3.83	13.6	3.12		
Loperamide [89]						9.3	5	35.9			
Cisapride [89]			1.56		1.93	7.32	17.1	10.8			
Ritanserin [89]			2.2			10.5	18.6	24	3.02		
Prucalopride [89]			0.43		4.3	17.6	8.77	10.6	4.57		
Sabeluzole [89]	8.41	1.83	5.37		2.45	10.4	37.7	29.2	0.83	2.95	5.48
Lubeluzole [89]			4.13			9.9	27.7	18.1	2.04		
Laniquidar [89]			2.86		5.82	12	16.8	38.7	7.07		
Acebutolol-R [80]	1.1	0.06	0.48	22.43	5.71	23.58	31.48	10.31	4.97	3.01	
Acebutolol-S [80]	0.79	0.04	0.36	91.25	4.3	32.7	24.89	6.14	4.45	2.47	
Betaxolol-R [80]	2.95	13.2	12.93	40.23	23.59	58.3	130.91	203.52	13.78	6.52	
Betaxolol-S [80]	2.86	12.85	13.01	37.8	21.52	54.54	108	182.52	13.55	6.05	
Biperiden [2]	67.64	2.28	7.95	12.92	8.04	12.13		86.31	3.69	4.7	
Bisoprolol-R [80]	1.03	4.88	1.64	26.52	6.49	24.91	22.78	41.82	5.4	2.18	
Bisoprolol-S [80]	1.02	4.43	1.79	25.67	6.69	24.82	22.95	41.99	5.23	2.21	
Carvedilol-R [80]	0.8				1.94	1.92	4.52	34	0.81		
Carvedilol-S [80]	1.9				7.42	7	11.77	75.6	1.6		
Clozapine [2]			20								
Cotinine [2]	0.08		0.42	0.64	0.51	0.99	0.64	0.63	0.67		
Diazepam [3]	23.54	5.45	2.13	7.06	5.56	4.15	13.44	5.89	2.77	4.23	
Haloperidol [2]		27.2	13.37	10.8	14.3			53.5	29	6.2	
Imipramine [53]	7.35		22.99	26.66	21.91	54.19	121.28	141.22	9.91	1.68	57.36
Inaperisone [82]	16		12		7.4	58	34	33	4.1	6.3	
Lidocaine [2]			3.24	3.12	2.73	17.21	11.51	3.8	1.68	2.58	4.79
Metoprolol-R [80]	1.04	5.18	6.48	12.96	6.89	26.56	40.04	25.56	5.66	3.19	
Metoprolol-S [80]	0.98	5.33	6.97	11.22	6.25	26.89	44.59	26.57	5.57	2.92	
Morphine [83,107]						9.5	1.2		2.5		
Nicotine [2]	0.32		2.02	1.6	1.12	18.14	4.95	1.24	1.23	1.1	
Oxprenolol-R [80]	0.58	1.87	1.29	12.71	3.62	14.17	8.59	15.86	3.08	1.37	

Oxprenolol-S [80]	0.69	2.33	2.48	11.18	4.37	17.62	12.33	21.24	3.89	1.7	
Pentazocine [2]	2.5	5.4	4.3	4.7	5.4	20	2.3	27	5.9	4.7	
Pindolol-R [80]	0.88	2.71	5.1	26.01	13.87	47.4	14.36	33.58	8.08	2.86	
Pindolol-S [80]	0.62	2.29	5.17	18.32	9.27	29.79	7.24	30.32	7.28	2.74	
Procainamide [2]	0.13		2.47		2.48	6.38	3.19		4.38		
Promazine [2]			62.5								
Propranolol [2]			14	6.6	7.1	15.3	11.6	16.46	4.3		14.2
Propranolol-R [80]	0.65	1.39	6.51	6.27	3.86	6.19	5.56	24.24	1.89	1.09	
Propranolol-S [80]	2.41	6.73	35.69	23.11	15.75	35.31	29.34	131.7	9.4	5.21	
Pyridostigmine [2]					1.1	15.2	2.1		0.52		
Quinidine [86]			1.16	14.42	8.92	19.51	20.79	44.03	3.82		23.99
Theophyllin [2]			0.36					0.71	0.6		
Thioridazine [2]			1.4								
Timolol-S [80]	0.64	1	1.06	20.16	5.36	13.32	7.87	26.96	4.15	1.58	
Verapamil [6]					6	12.5		50	3.5		
Bromperidol [2]			24								
Fluphenazine [2]			30.8								
Ftorafur [9]	0.17		0.41	0.36	0.38	0.68	0.39	0.26	0.5	0.4	0.42
Medazepam [2]									2.2		
Neostigmine [2]									0.59		
N-Methylpentobarbital [2]									1.3		
Propofol [2]			8.2								
2,3-Dideoxyinosine [2]			0.46	0.51		6.86	0.77		0.69		0.96
Clobazam [2]									2.6		
Cyclosporin [98]	11.57	3.18	0.79	5.23	4.05	7.99	12.2	5.52	1.35	2.92	5.45
Digoxin [2]				5.91	1.65	2.07	15.19	2.09	1.4		
Ethoxybenzamide [2]	0.71		0.94	0.56	0.99	1.3		0.91	0.81	1.04	0.87
Chlordiazepoxide [9]	4.31		0.75	1.97	2.61	2.7	4.85		0.77	0.48	
Prazepam [2]									1.8		
Triazolam [3]	6.02			11.9		8.43	3.75		6.02	5.46	
Alfentanil [2]	1.89		0.13	1.18	0.55	0.82	1	0.78	0.31	0.18	0.73
Alprazolam [3]	1.08	0.95	1.88	1.67	1.69	3.68	8.39	3.15	2	2.96	
Flunitrazepam [9]	73.5	4.36	1.46	2.74	1.66	0.4	3.69	4.78	1.03		
Midazolam [3]	4.62	1.92	2.49	2.81	4.64	3.19	8.51	4.08	0.87	1.96	2.42
Sparfloxacin [92]	0.18			9.87	2.09	7.55	4.5	2.45	1.93	2.08	
Ceftazidime [102]	0.16			0.41	0.22	4.8	0.25	0.44	0.19	0.39	
Nalidixic acid [87]		0.29	0.22	0.49	0.49	0.54	0.58	0.33	0.36	0.35	
Enoxacin [87]		1.44			1.07	4.61	3.21	1.14	1.45	1.36	1.63

Lomefloxacin [87]	0.27	1.58	0.22	1.63	1.37	4.84	2.3	1.24	1.61	0.94	1.73
Ofloxacin [87]	0.19	1.42	0.24		1.78	6.39	2.04	1.36	1.72	1.19	1.93
Grepafloxacin [91]				6.06	5.19	15.01	11.46	20.23	3.54		
Norfloxacin [2]								1.34	0.92		
Pefloxacin [2]			0.16		2.36	4.13	5.34	1.94	2.41		3.42
Pipemidic acid [2]	0.34	2.02	0.13		0.89	7.41	4.61	1.03	1.05		1.35
Tetracycline [2]	1.1	8.11		3.75		4.05	4.7		1.62		

Bibliography

1. Peters SA. Physiologically-Based Pharmacokinetic (PBPK) Modeling and Simulations: Principles, Methods, and Applications in the Pharmaceutical Industry. Wiley, 2012.

2. Poulin P, Theil FP. A priori prediction of tissue:plasma partition coefficients of drugs to facilitate the use of physiologically-based pharmacokinetic models in drug discovery. J Pharm Sci 2000; 89(1):16-35.

3. Gueorguieva I, Nestorov IA, Murby S, et al. Development of a whole body physiologically based model to characterise the pharmacokinetics of benzodiazepines. 1: Estimation of rat tissue-plasma partition ratios. J Pharmacokinet Pharmacodyn 2004; 31(4):269-298.

4. Bjorkman S. Prediction of the volume of distribution of a drug: which tissue-plasma partition coefficients are needed? J Pharm Pharmacol 2002; 54(9):1237-1245.

5. Jansson R, Bredberg U, Ashton M. Prediction of drug tissue to plasma concentration ratios using a measured volume of distribution in combination with lipophilicity. J Pharm Sci 2008; 97(6):2324-2339.

6. Schmitt W. General approach for the calculation of tissue to plasma partition coefficients. Toxicol In Vitro 2008; 22(2):457-467.

7. Poulin P, Theil FP. Development of a novel method for predicting human volume of distribution at steady-state of basic drugs and comparative assessment with existing methods. J Pharm Sci 2009; 98(12):4941-4961.

8. Rodgers T, Leahy D, Rowland M. Physiologically based pharmacokinetic modeling 1: predicting the tissue distribution of moderate-to-strong bases. J Pharm Sci 2005; 94(6):1259-1276.

9. Rodgers T, Rowland M. Physiologically based pharmacokinetic modelling 2: predicting the tissue distribution of acids, very weak bases, neutrals and zwitterions. J Pharm Sci 2006; 95(6):1238-1257.

10. Poulin P, Krishnan K. A biologically-based algorithm for predicting human tissue: blood partition coefficients of organic chemicals. Hum Exp Toxicol 1995; 14(3):273-280.

11. Poulin P, Krishnan K. A tissue composition-based algorithm for predicting tissue:air partition coefficients of organic chemicals. Toxicol Appl Pharmacol 1996; 136(1):126-130.

12. Berezhkovskiy LM. Volume of distribution at steady state for a linear pharmacokinetic system with peripheral elimination. J Pharm Sci 2004; 93(6):1628-1640.

13. Peyret T, Poulin P, Krishnan K. A unified algorithm for predicting partition coefficients for PBPK modeling of drugs and environmental chemicals. Toxicol Appl Pharmacol 2010; 249(3):197-207.

14. Yata N, Toyoda T, Murakami T, et al. Phosphatidylserine as a Determinant for the Tissue Distribution of Weakly Basic Drugs in Rats. Pharmaceutical Research 1990; 7(10):1019-1025.

15. Poulin P, Ekins S, Theil FP. A hybrid approach to advancing quantitative prediction of tissue distribution of basic drugs in human. Toxicol Appl Pharmacol 2011; 250(2):194-212.

16. Poulin P, Schoenlein K, Theil FP. Prediction of adipose tissue: plasma partition coefficients for structurally unrelated drugs. J Pharm Sci 2001; 90(4):436-447.

17. Fichtl B, Kurz H. Binding of drugs to human muscle. European Journal of Clinical Pharmacology 1978; 14(5):335-340.

18. Arundel P. Modeling and control in biomedical systems. IFAC Symposium 3rd. 1997. Warwick, UK.
Ref Type: Generic

19. Yun YE, Edginton AN. Correlation-based prediction of tissue-to-plasma partition coefficients using readily available input parameters. yun. 43[0]. 2013. Xenobiotica.
Ref Type: Generic

20. Bailer AJ, Dankovic DA. An introduction to the use of physiologically based pharmacokinetic models in risk assessment. Stat Methods Med Res 1997; 6(4):341-358.

21. Edginton AN, Schmitt W, Willmann S. Development and evaluation of a generic physiologically based pharmacokinetic model for children. Clin Pharmacokinet 2006; 45(10):1013-1034.

22. Edginton AN, Willmann S. Physiology-based simulations of a pathological condition: prediction of pharmacokinetics in patients with liver cirrhosis. Clin Pharmacokinet 2008; 47(11):743-752.

23. Rowland YK, Jamei M, Yang J, et al. Physiologically based mechanistic modelling to predict complex drug-drug interactions involving simultaneous competitive and time-

dependent enzyme inhibition by parent compound and its metabolite in both liver and gut - the effect of diltiazem on the time-course of exposure to triazolam. Eur J Pharm Sci 2010; 39(5):298-309.

24. Zhao P, Zhang L, Grillo JA, et al. Applications of physiologically based pharmacokinetic (PBPK) modeling and simulation during regulatory review. Clin Pharmacol Ther 2011; 89(2):259-267.

25. Andersen ME. Physiological modelling of organic compounds. Ann Occup Hyg 1991; 35(3):309-321.

26. Andersen ME. Development of physiologically based pharmacokinetic and physiologically based pharmacodynamic models for applications in toxicology and risk assessment. Toxicol Lett 1995; 79(1-3):35-44.

27. Payne MP, Kenny LC. Comparison of models for the estimation of biological partition coefficients. J Toxicol Environ Health A 2002; 65(13):897-931.

28. Edginton AN, Theil FP, Schmitt W, et al. Whole body physiologically-based pharmacokinetic models: their use in clinical drug development. Expert Opin Drug Metab Toxicol 2008; 4(9):1143-1152.

29. Poulin P, Theil FP. Prediction of pharmacokinetics prior to in vivo studies. 1. Mechanism-based prediction of volume of distribution. J Pharm Sci 2002; 91(1):129-156.

30. Sawada Y, Hanano M, Sugiyama Y, et al. Prediction of the volumes of distribution of basic drugs in humans based on data from animals. J Pharmacokinet Biopharm 1984; 12(6):587-596.

31. Toutain PL, Bousquet-Melou A. Volumes of distribution. J Vet Pharmacol Ther 2004; 27(6):441-453.

32. Joshi G, Tremblay RT, Martin SA, et al. Partition coefficients for nonane and its isomers in the rat. Toxicol Mech Methods 2010; 20(9):594-599.

33. Graham H, Walker M, Jones O, et al. Comparison of in-vivo and in-silico methods used for prediction of tissue: plasma partition coefficients in rat. J Pharm Pharmacol 2012; 64(3):383-396.

34. Jones RD, Jones HM, Rowland M, et al. PhRMA CPCDC initiative on predictive models of human pharmacokinetics, part 2: Comparative assessment of prediction methods of human volume of distribution. J Pharm Sci 2011: 100(10): 4074-4089.

35. Panchagnula R, Thomas NS. Biopharmaceutics and pharmacokinetics in drug research. Int J Pharm 2000; 201(2):131-150.

36. Toon S, Rowland M. Structure-pharmacokinetic relationships among the barbiturates in the rat. J Pharmacol Exp Ther 1983; 225(3):752-763.

37. Civelek VN, Hamilton JA, Tornheim K, et al. Intracellular pH in adipocytes: effects of free fatty acid diffusion across the plasma membrane, lipolytic agonists, and insulin. Proc Natl Acad Sci U S A 1996; 93(19):10139-10144.

38. Harrison DK, Walker WF. Micro-electrode measurement of skin pH in humans during ischaemia, hypoxia and local hypothermia. J Physiol 1979; 291:339-350.

39. Malan A, Rodeau JL, Daull F. Intracellular pH in hibernation and respiratory acidosis in the European hamster. J Comp Physiol B 1985; 156(2):251-258.

40. Schanker LS, Less MJ. Lung pH and pulmonary absorption of nonvolatile drugs in the rat. Drug Metab Dispos 1977; 5(2):174-178.

41. Waddell WJ, Bates RG. Intracellular pH. Physiol Rev 1969; 49(2):285-329.

42. Wood SC, Schaefer KE. Regulation of intracellular pH in lungs and other tissues during hypercapnia. J Appl Physiol 1978; 45(1):115-118.

43. Rothe KF, Heisler N. Correction of metabolic alkalosis by HCl and acetazolamide: effects on extracellular and intracellular acid-base status in rats in vivo. Acta Anaesthesiol Scand 1986; 30(7):566-570.

44. Murakami T, Yumoto R. Role of phosphatidylserine binding in tissue distribution of amine-containing basic compounds. Expert Opin Drug Metab Toxicol 2011; 7(3):353-364.

45. Obach RS, Lombardo F, Waters NJ. Trend analysis of a database of intravenous pharmacokinetic parameters in humans for 670 drug compounds. Drug Metab Dispos 2008; 36(7):1385-1405.

46. Wilkinson GR. Plasma and tissue binding considerations in drug disposition. Drug Metab Rev 1983; 14(3):427-465.

47. Gaulton A, Bellis LJ, Bento AP, et al. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 2012; 40(Database issue):D1100-D1107.

48. R Development Core Team. R: A Language and Environment for Statistical Computing. 2008. Vienna, Austria, R Foundation for Statistical Computing. Ref Type: Generic

49. Akaike H. A new look at the statistical model identification. Automatic Control, IEEE Transactions 1974; 19(6):716-723.

50. Montgomery DC, Peck EA, Vining GG. Introduction to linear regression analysis. 4 ed. Wiley, 2006.

51. Daniel WA, Wojcikowski J. Contribution of lysosomal trapping to the total tissue uptake of psychotropic drugs. Pharmacol Toxicol 1997; 80(2):62-68.

52. Siebert GA, Hung DY, Chang P, et al. Ion-trapping, microsomal binding, and unbound drug distribution in the hepatic retention of basic drugs. J Pharmacol Exp Ther 2004; 308(1):228-235.

53. Ishizaki J, Yokogawa K, Hirano M, et al. Contribution of lysosomes to the subcellular distribution of basic drugs in the rat liver. Pharm Res 1996; 13(6):902-906.

54. Patel MM, Goyal BR, Bhadada SV, et al. Getting into the brain: approaches to enhance brain drug delivery. CNS Drugs 2009; 23(1):35-58.

55. Lin JH, Lu AY. Role of pharmacokinetics and metabolism in drug discovery and development. Pharmacol Rev 1997; 49(4):403-449.

56. Lin JH, Yamazaki M. Role of P-glycoprotein in pharmacokinetics: clinical implications. Clin Pharmacokinet 2003; 42(1):59-98.

57. Dobson PD, Kell DB. Carrier-mediated cellular uptake of pharmaceutical drugs: an exception or the rule? Nat Rev Drug Discov 2008; 7(3):205-220.

58. Large CH, Bison S, Sartori I, et al. The efficacy of sodium channel blockers to prevent phencyclidine-induced cognitive dysfunction in the rat: potential for novel treatments for schizophrenia. J Pharmacol Exp Ther 2011; 338(1):100-113.

59. Liu W, Zi M, Naumann R, et al. Pak1 as a novel therapeutic target for antihypertrophic treatment in the heart. Circulation 2011; 124(24):2702-2715.

60. Japkowicz N, Shah M. Evaluating Learning Algorithms: A Classification Perspective. Cambridge University Press, 2011.

61. Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychol Methods 2009; 14(4):323-348.

62. Therneau TM, Atkinson EJ. An introduction to recursive partitioning using the RPART routines. 1997. Technical Report 61, Section of Biostatistics, Mayo Clinic, Rochester. URL http://www. mayo. edu/hsr/techrpt/61. pdf.
Ref Type: Report

63. Breiman L, Friedman JH, Olshen RA, et al. Classification and Regression Trees. Belmont, California: Wadsworth. 1984. Inc.
Ref Type: Generic

64. Venables WN, Ripley BD. Modern applied statistics with S. Springer, 2002.

65. Therneau TM, Atkinson B, Ripley B. Rpart: recursive partitioning. R package version 2010; 3:1-46.

66. Efron B. Bootstrap methods: another look at the jackknife. The annals of Statistics 1979; 7(1):1-26.

67. Breiman L. Random forests. Machine learning 2001; 45(1):5-32.

68. Touw WG, Bayjanov JR, Overmars L, et al. Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? Brief Bioinform 2012.

69. Liaw A, Wiener M. Classification and Regression by randomForest. R news 2002; 2(3):18-22.

70. Svetnik V, Liaw A, Tong C, et al. Application of BreimanΓÇÖs random forest to modeling structure-activity relationships of pharmaceutical molecules. Multiple Classifier Systems 2004;334-343.

71. Breiman L. Bagging predictors. Machine learning 1996; 24(2):123-140.

72. Peters A, Hothorn T, Lausen B. ipred: Improved predictors. R news 2002; 2(2):33-36.

73. Rowland M, Tozer T. Clinical pharmacokinetics/pharmacodynamics. Lippincott Williams and Wilkins, 2005.

74. Paixão P, Gouveia LsF, Morais JA. Prediction of drug distribution within blood. european journal of pharmaceutical sciences 2009; 36(4):544-554.

75. Leo A, Hansch C, Elkins D. Partition coefficients and their uses. Chem Rev 1971; 71(6):525-616.

76. Small H, Gardner I, Jones HM, et al. Measurement of binding of basic drugs to acidic phospholipids using surface plasmon resonance and incorporation of the data into mechanistic tissue composition equations to predict steady-state volume of distribution. Drug Metab Dispos 2011; 39(10):1789-1793.

77. Rodgers T, Rowland M. Mechanistic approaches to volume of distribution predictions: understanding the processes. Pharmaceutical Research 2007; 24(5):918-933.

78. Haddad S, Poulin P, Krishnan K. Relative lipid content as the sole mechanistic determinant of the adipose tissue: blood partition coefficients of highly lipophilic organic chemicals. Chemosphere 2000; 40(8):839-843.

79. Poulin P, Haddad S. Advancing prediction of tissue distribution and volume of distribution of highly lipophilic compounds from a simplified tissue-composition-based model as a mechanistic animal alternative method. J Pharm Sci 2012; 101(6):2250-2261.

80. Rodgers T, Leahy D, Rowland M. Tissue distribution of basic drugs: accounting for enantiomeric, compound and regional differences amongst beta-blocking drugs in rat. J Pharm Sci 2005; 94(6):1237-1248.

81. Vaille A, Balansard G, Jadot G. Effects of a subacute treatment in rats by a fresh cola extract on EEG and pharmacokinetics. Pharmacol Biochem Behav 1993; 45(4):791-796.

82. Nagata O, Murata M, Kato H, et al. Physiological pharmacokinetics of a new muscle-relaxant, inaperisone, combined with its pharmacological effect on blood flow rate. Drug Metab Dispos 1990; 18(6):902-910.

83. Gabrielsson JL, Paalzow LK. A physiological pharmacokinetic model for morphine disposition in the pregnant rat. J Pharmacokinet Biopharm 1983; 11(2):147-163.

84. Plowchalk DR, Andersen ME, deBethizy JD. A physiologically based pharmacokinetic model for nicotine disposition in the Sprague-Dawley rat. Toxicol Appl Pharmacol 1992; 116(2):177-188.

85. Hanada K, Akimoto S, Mitsui K, et al. Enantioselective tissue distribution of the basic drugs disopyramide, flecainide and verapamil in rats: role of plasma protein and tissue phosphatidylserine binding. Pharm Res 1998; 15(8):1250-1256.

86. Mansor SM, Ward SA, Edwards G. The effect of fever on quinine and quinidine disposition in the rat. J Pharm Pharmacol 1991; 43(10):705-708.

87. Okezaki E, Terasaki T, Nakamura M, et al. Structure-tissue distribution relationship based on physiological pharmacokinetics for NY-198, a new antimicrobial agent, and the related pyridonecarboxylic acids. Drug Metab Dispos 1988; 16(6):865-874.

88. Olanoff L, Anderson JM. Controlled release of tetracycline II: Development of an in vivo flow-limited pharmacokinetic model. J Pharm Sci 1979; 68(9):1151-1155.

89. De Buck SS, Sinha VK, Fenu LA, et al. The prediction of drug metabolism, tissue distribution, and bioavailability of 50 structurally diverse compounds in rat using mechanism-based absorption, distribution, and metabolism prediction tools. Drug Metab Dispos 2007; 35(4):649-659.

90. De Buck SS, Sinha VK, Fenu LA, et al. Prediction of human pharmacokinetics using physiologically based modeling: a retrospective analysis of 26 clinically tested drugs. Drug Metab Dispos 2007; 35(10):1766-1780.

91. Nakajima Y, Hattori K, Shinsei M, et al. Physiologically-based pharmacokinetic analysis of grepafloxacin. Biol Pharm Bull 2000; 23(9):1077-1083.

92. Hayakawa H, Takagi K, Takano YF, et al. Determinant of the distribution volume at steady state for novel quinolone pazufloxacin in rats. J Pharm Pharmacol 2002; 54(9):1229-1236.

93. Tsuji A, Miyamoto E, Terasaki T, et al. Physiological pharmacokinetics of beta-lactam antibiotics: penicillin V distribution and elimination after intravenous administration in rats. J Pharm Pharmacol 1979; 31(2):116-119.

94. Yoshikawa T, Sugiyama Y, Sawada Y, et al. Effect of pregnancy on tissue distribution of salicylate in rats. Drug Metab Dispos 1984; 12(4):500-505.

95. Ballard P, Leahy DE, Rowland M. Prediction of in vivo tissue distribution from in vitro data. 3. Correlation between in vitro and in vivo tissue distribution of a homologous series of nine 5-n-alkyl-5-ethyl barbituric acids. Pharm Res 2003; 20(6):864-872.

96. Igari Y, Sugiyama Y, Sawada Y, et al. Prediction of diazepam disposition in the rat and man by a physiologically based pharmacokinetic model. J Pharmacokinet Biopharm 1983; 11(6):577-593.

97. Itoh T, Sawada Y, Lin TH, et al. Kinetic analysis of phenytoin disposition in rats with experimental renal and hepatic diseases. J Pharmacobiodyn 1988; 11(5):289-308.

98. Bernareggi A, Rowland M. Physiologic modeling of cyclosporin kinetics in rat and man. J Pharmacokinet Biopharm 1991; 19(1):21-50.

99. Bjorkman S, Stanski DR, Verotta D, et al. Comparative tissue concentration profiles of fentanyl and alfentanil in humans predicted from tissue/blood partition data obtained in rats. Anesthesiology 1990; 72(5):865-873.

100. Meno-Tetang GM, Li H, Mis S, et al. Physiologically based pharmacokinetic modeling of FTY720 (2-amino-2[2-(-4-octylphenyl)ethyl]propane-1,3-diol hydrochloride) in rats after oral and intravenous doses. Drug Metab Dispos 2006; 34(9):1480-1487.

101. Sugita O, Sawada Y, Sugiyama Y, et al. Physiologically based pharmacokinetics of drug-drug interaction: a study of tolbutamide-sulfonamide interaction in rats. J Pharmacokinet Biopharm 1982; 10(3):297-316.

102. Granero L, Santiago M, Cano J, et al. Analysis of ceftriaxone and ceftazidime distribution in cerebrospinal fluid of and cerebral extracellular space in awake rats by in vivo microdialysis. Antimicrob Agents Chemother 1995; 39(12):2728-2731.

103. Brocks DR, Jamali F. Enantioselective pharmacokinetics of etodolac in the rat: tissue distribution, tissue binding, and in vitro metabolism. J Pharm Sci 1991; 80(11):1058-1061.

104. Kawahara M, Nanbo T, Tsuji A. Physiologically based pharmacokinetic prediction of pΓÇÉphenylbenzoic acid disposition in the pregnant rat. Biopharmaceutics & drug disposition 1998; 19(7):445-453.

105. La Rosa C, Mather LE, Morgan DJ. Pethidine binding in plasma: effects of methodological variables. British journal of clinical pharmacology 1984; 17(4):411-415.

106. La Rosa C, Morgan DJ, Mather LE. Pethidine binding in whole blood: methodology and clinical significance. British journal of clinical pharmacology 1984; 17(4):405-409.

107. Bhargava HN, Villar VM, Rahmani NH, et al. Distribution of morphine in brain regions, spinal cord and serum following intravenous injection to morphine tolerant rats. Brain Res 1992; 595(2):228-235.