The Application of Quantitative Structure Activity Relationship Models to the Method Development of Countercurrent Chromatography

A thesis submitted for the degree of Doctor of Philosophy

By Siân Catherine Marsden-Jones

Advanced Bioprocessing Centre,
Institute of Environment, Health and Societies, Brunel University

August 2015

I would like to dedicate this work to my parents and my partner for their support, encouragement and patience.

Abstract

A fundamental challenge for liquid-liquid separation techniques such as countercurrent chromatography (CCC) and centrifugal partition chromatography (CPC), is the swift, efficient selection of the two phase solvent system containing more than two solvents, for the purification of pharmaceuticals and other molecules. A purely computational model that could predict the optimal solvent systems for separation using just molecular structure would be ideal for this task. The experimental value being predicted is the partition coefficient (K_d), which is the concentration of the compound in one phase divided by the concentration in the other. Using this approach, Quantitative Structure Activity Relationship (QSAR) models have been developed to predict the partitioning of compounds in two phase systems from the molecular structure of the compound using molecular descriptors. A K_d value in the range of 0.5 to 2 will give optimal separation. Molecular descriptors are varied, examples include logP values, hydrogen bond donor values and the number of oxygen atoms.

This work describes how the QSAR models were developed and tested. A dataset of experimental logK_d values for 54 compounds in six different combinations of four solvents (heptane, ethyl acetate, methanol and water) was used to train the QSAR models. A set of 196 possible molecular descriptors was generated for the 54 compounds and a partial least squares regression was used to identify which of these was significant in the relationship between logK_d and molecular structure. The resulting models were used to predict the logK_d values of four test compounds that had not been used to build the QSAR models. When these predictions were compared to the experimental logK_d values, the root mean squared error for four of the six models was less than 0.5 and less than 0.7 for the remaining two.

These models were used to successfully separate a range of structurally diverse pharmaceutical compounds by predicting the best solvent systems to carry out the separation on the CCC/CPC using nothing but their molecular structure.

Table of Contents

Table of Equations

Table of Figures

Table of Tables

Acknowledgements

Firstly, I would like to acknowledge AstraZeneca and the EPSRC for funding this work and AstraZeneca for allowing me the use of multiple software packages. I would like to thank my industrial supervisors, Neil Sumner and Nicola Colclough, for all the advice and assistance they have given me during my trips to Alderley Park and the many telecom and Webex meetings, despite the technical difficulties often encountered!

My Brunel supervisors, Svetlana Ignatova and Ian Garrard, have been wonderful and are a huge part of why I have had such an enjoyable PhD experience. Thank you for your help, support, guidance and expertise. You have never failed to drop everything to help with problems and queries, whilst allowing me the space to become into an independent researcher. This section would not complete without an acknowledgment for Svetlana's generous use of "her spot" during lunch times. I hope she enjoys having it back. I also hope her new glasses allow her to wear one pair at a time! In addition, although Ian's attempt at my suggestion for microwave cooking didn't end so well, I hope he will persevere in my absence without causing too much smoke damage to his house!

I would also like to thank Peter Hewitson for all the help he provided in the laboratory and thank him for his patience on the many occasions when he was the only thing standing between me and a meltdown, when the HPLC had broken for the thousandth time. Additionally, I would like to acknowledge Jenny Kume who has been a wonderful support, even before I had arrived at Brunel. I am yet to find a question that she could not answer or a problem that she could not solve. I would also like to extend my thanks to Ian Sutherland for his guidance and support, and for the occasional lift to and from Bristol. Thanks is also due to Jon Huddleston for challenging me and for introducing me to Michael Abraham who provide invaluable insight at the start of my PhD.

A very special mention must go to the students within former Brunel Institute for Bioengineering (BIB) for their part in making these three years so fantastic. I hope the friendships I have formed here will continue for many years to come. I would especially like to thank Piyali Basu for her support, the laughter, the walks and for being my housemate but mostly for being an incredible listener. I'm sure many more trips to Costa await!

My Mum, Dad, David and Granny have provided enormous amounts of support throughout these three years. From Dad being my de facto landlord, fixing leaks and waiting in for delivery men, to Mum proof-reading the final version of my thesis, I'm so grateful for everything you have done for me to make my life so much easier.

Finally, I would like to thank Adrian van Arkel, as without his unconditional love and support, none of this would have been possible. Thank you for all your advice and encouragement, and for the hundreds (if not thousands) of trips along the M4. I couldn't have done it without you!

Abbreviations

Software

Absolv	Prediction of Abraham's parameters
ACDlabs	Advanced Chemistry Development Labs
CoMFA	Comparative Molecular Field Analysis
COSMO-RS	COnductor-like Screening MOdel for Real Solvents
DoE	Design of Experiments
GA	Generic Algorithm
LIBSVM	A Library for Support Vector Machines
MATLAB	Matrix Laboratory
NRTL-SAC	Non-Random Two Liquid – Segment Activity Coefficient
PrologD	Predicting distribution coefficients
RF 4.6-6	R's Random Forest algorithm 4.6-6
SLIPPER	Solubility LIPophilicity PERmeability
SPARC	Scalable Process Architecture

Solvent Systems

ChMWat	Chloroform, Methanol, Water
EBuWat	Ethyl Acetate/n-butanol/water
HEMWat	Heptane, Ethyl Acetate, Methanol and Water
HIAW	Heptane, Isopropyl acetate, Acetone, Water
HMAW	Heptane, Methanol, Acetone and Water

HTAW	Heptane, Toluene, Acetone and Water
H*ter*AcWat	Hexane/t-butylmethylether/acetonitrile/water
MTBE	tert-butyl methyl ether
terAcWat	t-buymethylether/acetonitrile/water

Descriptors

A	Abraham Parameter; Hydrogen Bonding Acidity
α	Abraham hydrogen bonding acidity coefficient
ACD LogP	Calculated LogP value from the software “Advanced Chemistry Development”
ACDLogD pH7.4	ACDLogD74 is calculated as the octanol/water distribution coefficient at pH 7.4 from “Advanced Chemistry Development”
ACDLogD pH6.5	ACDLogD65 is calculated as the octanol/water distribution coefficient at pH 6.5 from “Advanced Chemistry Development”
B	Abraham Parameter; Hydrogen Bonding Basicity
β	Hydrogen Bonding Basicity coefficient
ClogP	Predicted value of the octanol/water partition coefficient.
E	Abraham Parameter; Excessive Molar Refraction
e	coefficient of molar refractivity term
GClogP	Octanol/water partition coefficient based on Ghose/Crippen atom types
HBA	Lipinski number of hydrogen bond acceptors
HBD	Lipinski number of hydrogen bond donors
HBD Selma	Number of hydrogen bond donors

HBsumTotal	Sum of donor and acceptor free energies according to Raevsky
logD	Distribution coefficient
logP	Octanol/water partition coefficient
logPalk	Partition coefficient between water and alkanes
MW	Molecular Weight
NNlogP	Octanol/water partition coefficient using a neural network approach based on Ghose/Crippen atom types
NPSA	Non-Polar Surface Area
RingCount	Number of rings
RotBond	Number of non-terminal flexible bonds
PSA	Polar Surface Area
S	Abraham Parameter; Polarity/ Polarisability
s	Abraham Parameter; Polarity/ Polarisability coefficient
V	Abraham Parameter; McGowan volume
v	Abraham Parameter; McGowan volume coefficient
VOL	Gaussian molecular volume

Constants

c	Speed of light (2.99×10^{8} m s^{-1})
ε_o	Relative permittivity of free space (8.85×10^{-12} m^{-3} kg^{-1} s^{4} A^{2})
E_s	Taft's steric constant
g	Gravitational field (9.81 m s^{-2})

h	Planck's constant (6.626×10^{-34} m^2kgs^{-1})
N_A	Avogadro's constant (6.022 x 1023 mol^{-1})
R	Molar gas constant (8.314 m^2 kg s^2 K^1 mol^{-1})

Chemicals

DCM	Dichloromethane
DMSO	Dimethyl sulfoxide
PEG1000	Polyethylene Glycol
TMS	Tetramethylsilane

Mathematical methods

CV	Cross Validation
FB-QSAR	Fragment-based QSAR
LFER	Linear Free Energy Relationships
LOO	Regression method; Leave-One-Out
LSER	Linear Solvation Energy Relationships
LSO	Regression method; Leave-Some-Out
MLR	Multiple Linear Regression
MQSM	Molecular Quantum Similarity Measures
MVA	Multiple Variant Analysis
NN	Neural Networks
p-value	Probability value

PCA	Principal Component Analysis
PLS	Partial Least Squares
PMO	Pertubational Molecular Orbital
PRESS	Predicted residual sum of squares
Q^2	Predictive squared correlation coefficient
QSAR	Quantitative Structure Activity Relationships
R^2	Coefficient of determination
RBF	Radial Basis Function
RF	Random Forest
RMSE	Root Mean Square Error
SARs	Structure activity relationship
SVM	Support Vector Machine
VIF	Variance Inflation Factor
UNIFAC	Universal quasichemical functional-group activity coefficients

Liquid-liquid separation techniques

CCC	Countercurrent Chromatography
CPC	Centrifugal Partition Chromatography

Other Practical Analytical Techniques

DAD	Diode Array Detector
ELSD	Evaporating Light Scattering Detector

GC	Gas Chromatography
HPLC	High Performance Liquid Chromatography
MS	Mass spectrometry
TLC	Thin Layer Chromatography

Miscellaneous

$[solute]_{octanol}$	Concentration of solute dissolve in octanol
$[solute]_{water}$	Concentration of solute dissolved in the water
A	Eddy-diffusion parameter
A_M	Area of adsorbent occupied by 1 mole of mobile phase
ATPS	Aqueous Two Phase Systems
B	Diffusion Coefficient
BBB	Blood Brain Barrier
C	Mass transfer coefficient
C	Molar Concentration
C_m	Concentration in the mobile phase
C_s	Concentration in the stationary phase
E	Electric Field Strength
EDG	Electron Donating Group
E_T(solvent)	Molar electronic transition energy of solvent
E_T(TMS)	Molar electronic transition energy of solvent of tetramethylsilane
E_T(water)	Molar electronic transition energy of water

EWG	Electron Withdrawing Group
ΔG	Gibbs Free Energy
ΔG_0	Variation of the free adsorption energy of 1 mole of mobile phase
G_{ij}	Group Contribution
H	Plate height
H_0	Null hypothesis
K	Equilibrium constant for substituted reactant
K_0	Equilibrium constant for non-substituted reactant
K_d	Partition Coefficient
L	Length of column
N	Number of theoretical plates
p	Dipole Moment
p_{ak}	Loading
R_F	Retention Factor
R_s	Resolution
S_F	Stationary Phase Retention
SMILES	Simplified Molecular-Input Line-Entry System
SP	Some Property
T	Absolute temperature (Kelvin)
T	Torque (sum of moments of forces not acting along the same line)
t_a	X-scores
u	Linear velocity

u_a	Y-scores
V_c	Total volume of column
V_d	Dead volume
V_i	Cavity volume (solute)
V_j	Cavity volume (solvent)
V_m	Volume of the mobile phase
V_R	Retention Volume
V_s	Volume of the stationary phase
V_{max}	Frequency of the maximum frequency of the longest wavelength intramolecular charge-transfer π-π* absorption
w_b	Width of peak base
w_D	Width of peak
X_n	Mole fraction
X_{ij}	Structural feature
α	Separation factor
γ	Activity Coefficients
γ	Interfacial tension between two liquid phases
δ	Susceptibility the reaction to steric effects of substituents
$\Delta_{vap}E_m$	Molecular energy of vaporisation
ε_0	Snyder polarity
ε_R	Relative permittivity
λ_{cap}	Capillary wavelength

λ_{max}	Wavelength of the maximum of the longest wavelength
μ	Biology activity of the reference compound
$\nu_{c\text{-}C6H12}$	Maximum frequency of the solvatochromatic adsorption band in cyclohexane
ν_{DMSO}	Maximum frequency of the solvatochromatic adsorption band in DMSO
ν_S	Maximum frequencies of the solvatochromatic adsorption band of 4-methylnitrobenzene in solvent, S
π	Lipophilicity
π_S^*	Kamlet-Taft polarity in solvent, S
$\Delta\rho$	Density difference between the phases
ρ	Reaction constant
ρ^*	Susceptibility of a reaction to the electronic nature of substituents
σ	Electronic parameter
σ^*	Polar substituent constant
Φj	Physiochemical parameters

1. Introduction and Literature Review

Chromatography is the process of separating a mixture into its constituent parts. This is usually done by dissolving the mixture to be separated in a mobile phase which is passed over a stationary phase. The rate at which the individual components of the mixture partition between the mobile and stationary phases differs depending on a compounds' molecular structure. Components that spend more time in the mobile phase will travel faster than those with a greater affinity to the stationary phase. This will lead to separation. The stationary phase can be bound to a solid matrix (e.g. High Performance Liquid Chromatography) or can be a free flowing liquid that is immiscible with the mobile phase (e.g. Countercurrent Chromatography). Using a solid state stationary phase only allows the analyte to interact with a thin layer of stationary phase on the solid surface, whereas using a liquid stationary phase increases the loading capacity of the column as the analytes can occupy the whole volume of the liquid stationary phase (Berthod, et al., 2009). Solid-liquid separation techniques have other disadvantages that can be overcome by using liquid-liquid chromatography, including the limited loss of compound as it can be recovered from the liquid phases using gravimetric methods. Some liquid-liquid separation techniques can also be used to separate crude samples as there is no risk of blocking an expensive column. This lack of an expensive column also significantly decreases the running costs of liquid-liquid chromatography. Liquid-liquid chromatography has been shown to have many useful applications and is reproducible and scalable. A hurdle for the mainstream acceptance of the technique by industry is the lack of a predictive tool for fast, efficient solvent system selection. Such a tool has the potential to increase the technology's capability to meet all customer demands and allow automation of the technique.

1.1. Aims and Objectives

- To determine a protocol for accurate and reproducible experimental measurements of partition coefficient (K_d) values. This will be achieved through investigating the physical factors which affect the measurement of K_d to try to combat the wide variation from laboratory to laboratory of experimental K_d values for the same compounds. This will ensure the computational model is trained with data that has the smallest amount of experimental error possible.

- To develop Quantitative Structure Activity Relationship (QSAR) models between the molecular structure of a compound and its K_d value. This would allow the production of a model that can suggest suitable solvent systems tailored specifically to the compound under investigation.

1.2. Liquid-Liquid Partition Chromatography

1.2.1. Countercurrent Chromatography

Countercurrent Chromatography (CCC) is a form of liquid-liquid chromatography that was invented in 1966 by Ito (Ito, et al., 1966). It combines the principles of partition chromatography and liquid-liquid extraction. The separation occurs due to the partitioning of target compounds between two liquid phases of a solvent system, comprising of two or more immiscible solvents. A CCC column is a long tube (traditionally made from Teflon or stainless steel) wound around a drum, which rotates around its own axis and simultaneously revolves around the central axis of the device in planetary motion (Figure 1), creating a fluctuating g-field (Sutherland, et al., 2009). It is a hydrodynamic technique where the stationary liquid phase is retained in the rotating column by a combination of centrifugal force and the pumping effect of the mobile liquid phase (Ito, et al., 1970). However, as soon as the flow of the mobile phase is stopped, both phases move to the opposite ends of the CCC column: the upper phase moves to the “head” of the column and lower phase to the “tail”.

The main feature of CCC as a chromatographic technique is the absence of any solid support for the liquid stationary phase. In turn, this determines the main advantages of the technology, such as a complete recovery of a sample, high loading capacity, the ability to run the column in both normal and reversed elution modes and easy switching between solvent systems (Sumner, 2011). The main feature of CCC as a liquid extraction method is the continuous mode of separation. The planetary motion creates mixing and settling zones equivalent to thousands of liquid-liquid extraction steps done in a sequence. Since its invention, the CCC technology has undergone tremendous development resulting in a variety of column designs to increase separation efficiency. However, the main element of each CCC machine, the flying leads, remains the same. These are a pair of tubes connecting the column with ancillary equipment.

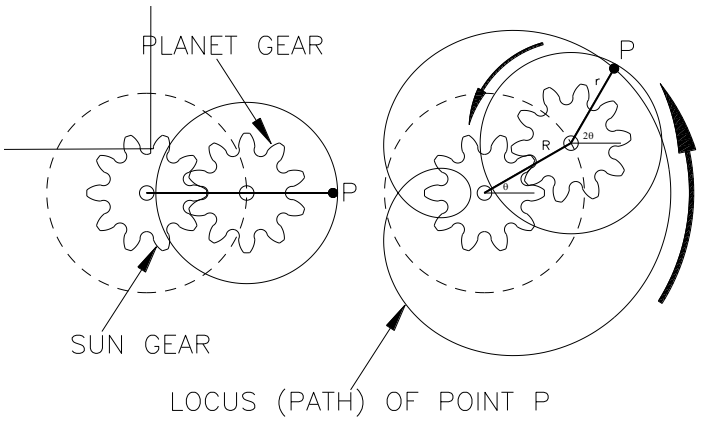


Figure 1 - The planetary motion of the J-type centrifuge (Wood, 2002)

Being a separation technique, CCC can be coupled with various on-line detectors (Michel, et al., 2013). The most common are a UV detector and a diode array detector (DAD) as both are non-destructive. If compounds do not have chromophores, CCC can also be used in conjunction with evaporative light scattering detector (ELSD). Since this technique involves the evaporation of mobile phase, the latter must be reasonably volatile and should not contain additives like salts. Therefore, it is not appropriate for all solvent systems. It should also be noted that ELSD is a destructive detector and should be used in conjunction with a split valve to avoid losing fractions. It is possible to have both detectors, DAD and ELSD, connected one after another in cases when a sample is a complex mixture with UV and non-UV visible compounds. A further option is to connect CCC machines to a mass spectrometer (MS). This approach is often used for screening purposes and for the fingerprinting of herbal extracts.

1.2.2. Centrifugal Partition Chromatography

In 1982 Japanese company, Sanki Engineering (Murayama, et al., 1982), introduced another type of CCC instrument with a new name to ensure it was distinct from CCC. The sister technology is well known as Centrifugal Partition Chromatography (CPC).

CPC involves a series of connected chambers linked by narrower channels (Figure 2). It is a hydrostatic technique with a single rotation axis and therefore, has a fixed gravitational field. When the flow of mobile phase is stopped, both stationary and mobile phase stay where they are. Mixing occurs as a cascade and is greatly dependent on a chambers' design (Sutherland, et al., 2005). The main difference

between CPC and CCC is the rotating seals that connect the column to the ancillary equipment and its much higher back pressure due to its narrow channels.

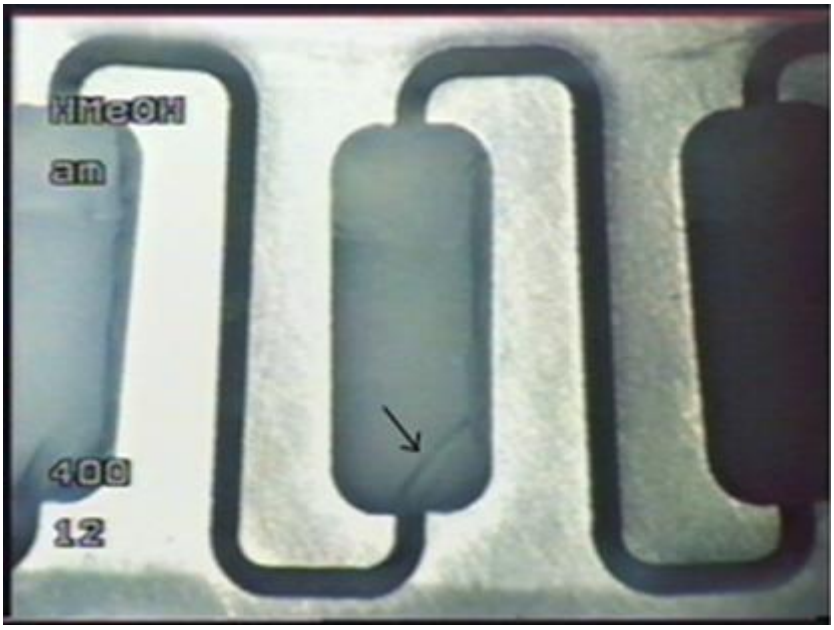


Figure 2 – Centrifugal Partition Chromatography

Although CCC and CPC columns have a completely different design, the separation occurs via the same principle; a compound partitioning between two liquid phases of a solvent system.

1.3. Principles of Liquid-Liquid Separation

The partitioning of a compound is governed by its partition coefficient (K_d) which was first developed by Berthelot and Jungfleisch in 1872 (Berthelot, et al., 1872). The partition coefficient (K_d) is defined in Equation 1.

$$K_d = \frac{C_s}{C_m}$$

Equation 1 – The partition coefficient where C_s is the concentration of a compound in the stationary phase and C_m is the concentration in the mobile phase.

The K_d value of a compound is dependent on the composition of the solvent system which is affected by temperature according to Nernst's distribution law (Margraff, et al., 1994). K_d values are independent of solute concentration at low concentrations. However, there can be a large variation in K_d values between ionised and non-ionised compounds. A pH change of one unit can change the K_d value of a compound by up

to ten times in comparison to the non-ionised K_d value (Conway, 1991). When a compound has a K_d value between 0.5 and 2 (Ren, et al., 2013), it is most likely that the compounds will have been in the column long enough to avoid co-elution, but not so long that separated peaks broaden and merge (Inoue, et al., 2012). A K_d value lower than 0.5 will lead to a loss of resolution and a K_d value higher than 2 will lead to a long retention time which will increase the run time required to carry out separation.

1.4. Experimental Measurement of Partition Coefficients

One of the methods of experimentally measuring the static K_d value of a compound is the shake flask method (Friesen, et al., 2005). This involves dissolving the compound in the solvent system which is the potential candidate to perform the separation and allowing the compound to partition. The solute is left in the system until dynamic equilibrium has been reached, then an aliquot of each of the upper and lower phases is sampled and the compounds' concentration is determined. This concentration is most commonly calculated using High Performance Liquid Chromatography (HPLC) but many other practical analytical techniques can be used. Equation 1 can then be used to calculate the K_d value of a compound.

A K_d value can be measured using Countercurrent Chromatography (CCC). As CCC is a liquid-liquid technique with no solid support for a stationary phase, the volume of the tubing used to make a CCC column is the volume of the column (Equation 2).

$$V_c = V_s + V_m$$

Equation 2 - V_c is the column volume, V_s is the volume of the stationary phase and V_m is the volume of the mobile phase.

The stationary phase retention (S_F) is an important factor in determining the success of a separation using CCC as higher stationary phase retention will lead to better resolution of the peaks.

$$S_F = \frac{V_s}{V_c}$$

Equation 3 - S_F is the phase volume ratio, V_c is the column volume and V_s is the volume of the stationary phase retained in the column.

It is recommended that a system with a minimum stationary phase retention of 50% is used. However, if the compound has a high K_d value then a stationary phase retention as low as 30% can be used for separation (Ito, 2005). S_F is dependent on the physicochemical properties of a solvent system used, column bore, temperature, spin speed, and mobile phase flow rate. The latter two are most commonly used to regulate the stationary phase retention. For conventional CCC single/multilayer columns, S_F can be increased by reducing the flow rate of the mobile phase (Ito, 2005) and increasing the spin speed. Changing these operating parameters will alter the K_d values of all of the compounds by either increasing or decreasing them. This results in a loss of resolution between peaks as they all converge around the K_d value of one (Berthod, et al., 2009). Separation efficiency is determined by the separation factor, α, which is the ratio of partition coefficients (Equation 4). When K_d is between 0.5 and 2 (Ren, et al., 2013) an α value of approximately 1.5 should provide the optimal balance between resolution and run time (V_R) (Ito, 2005).

$$\alpha = \frac{K_{d_2}}{K_{d_1}}$$

Equation 4 – The separation factor (α) where α is the separation factor and K_d is the partition coefficient ($K_{d_2} > K_{d_1}$).

Once compounds have been efficiently separated, the dynamic K_d value of the compounds can be calculated using the volume of the mobile and stationary phases along with the elution time of each compound (Equation 5).

$$V_R = V_m + V_s K_d$$

Equation 5 - V_R is the elution time of the analyte, V_m is the mobile phase volume, V_s is the stationary phase volume and K_d is the partition coefficient.

Combining Equation 2 and Equation 5 allows the prediction of the elution volume of a compound using just the stationary phase volume, the column volume and the partition coefficient (Equation 6). This does not take into account the dead volume of the system.

$$V_R = V_c + V_s(K_d - 1)$$

Equation 6 - V_R is the elution time, V_s is the stationary phase volume, V_c is the column volume, K_d the partition coefficient and S_F is the phase volume ratio.

Equation 7 can also be used for the calculation of the K_d value of a compound using the CCC chromatogram (Figure 3).

$$K_d = \frac{V_R - V_C}{V_s} + 1$$

Equation 7 – The partition coefficient (K_d) (Berthod, et al., 1991) where K_d the parition coefficient, V_R is the elution time, V_c is the column volume and V_s is the volume of the stationary phase.

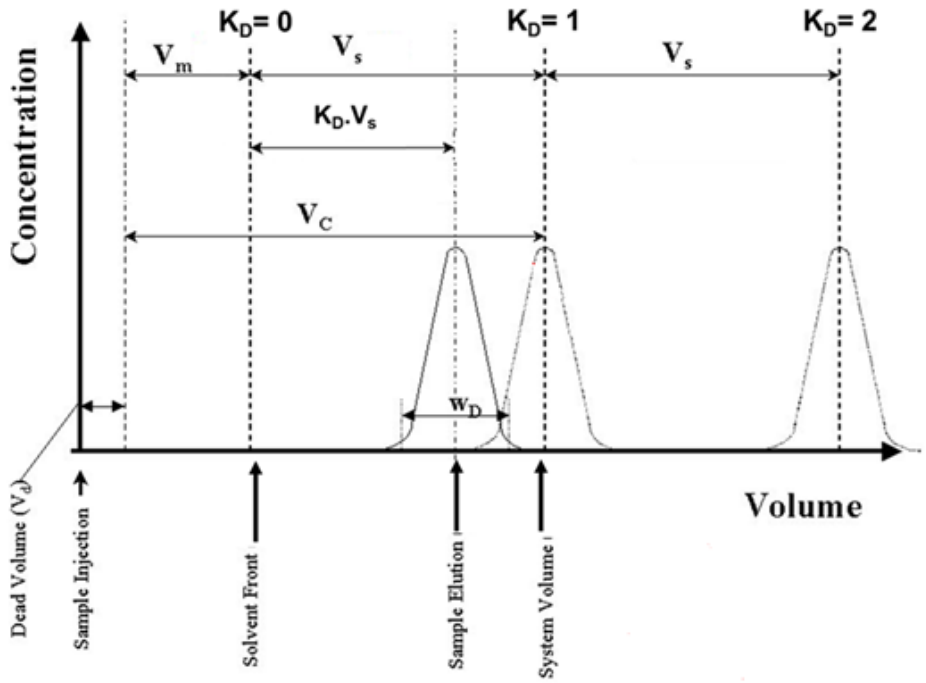


Figure 3 – A graphical representation of terms: the dead volume (V_d) is the volume that passes through the column before the sample hits the column. V_m is the volume of the mobile phase. V_S is the volume of the stationary phase. V_C is the total column volume which can be found by combining the mobile and stationary phase volumes. K_d is the partition coefficient and w_D is the peak width (Sutherland, 2002).

The liquid nature of both phases in CCC allows it to be run in both normal and reverse phase (Friesen, et al., 2005). Running the CCC machine in normal phase means using the more polar phase (traditionally lower aqueous phase) as the stationary phase and the less polar (traditionally upper organic phase) as the mobile phase. Running the CCC machine in reverse phase utilises the phases in the opposite roles. Using the

shake flask method to determine the K_d values of the compounds to be separated, will determine whether reverse or normal phase has the best likelihood of leading to optimal separation. If the upper phase is used as the stationary phase (reverse phase), the measured static K_d value from the shake flask method and the dynamic K_d value measured from the CCC chromatogram are equivalent. However, if the lower phase is the stationary phase (normal phase) then the static shake flask method K_d value will be equivalent to the reciprocal of the K_d value from the CCC chromatogram (Friesen, et al., 2009). When comparing the chromatograms produced in reverse phase and normal phase elution modes, they are the mirror image of each other with the K_d value of one being the central point. This allows compounds with very high K_d values in one elution mode, to be eluted by switching over the elution mode (the stationary and mobile phases) (Friesen, et al., 2009) or simply by extruding the column content (Berthod, et al., 2003).

1.5. Solvent Systems

The combination of solvents that form the mobile and stationary phases are called the "solvent system". By changing the solvents within a system or altering the percentage composition of the solvents, a large and diverse set of compounds can be separated. As long as the solvent system consists of more than one phase, any combination of solvents is a possible solvent system. The simplest examples are a combination of two solvents (hexane/water etc.). However, the separation of complex mixtures of analytes may require adding a third solvent that partitions between the two phases, and as a result improves peak width, the length and the efficiency of the separation (Foucault, 1995). There are also many examples of using four solvents to form a solvent system, e.g. heptane, toluene, acetone, water abbreviated to "HTAW" (Foucault, 1995). In this case, the two solvents, heptane and water, form the two phase structure and the other two solvents, toluene and acetone, act as modifiers partitioning between the organic and aqueous phases. By varying the composition of the solvents and their percentages, various "families" of solvent systems have been developed. Two widely used examples are the "ChMWat" family composed of chloroform, methanol and water (Oka, et al., 1991) and "HEMWat" composed of heptane, ethyl acetate, methanol and water. A review found that almost a third of the natural product protocols used the HEMWat system to perform the separation (Freisen, et al., 2015) and another found that 48% of the literature data used heptane based solvent systems

(Skalicka-Woźniak, et al., 2015). The second most commonly used solvent systems were butanol based systems which were described in 23% of the papers. One example of a solvent system family that contains butanol is ethyl acetate, n-butanol, and water (EBuWat). Ethers are another class of compounds which many solvent systems are based on, with two common examples being t-butylmethylether, acetonitrile, water (*ter*AcWat) and ether and hexane based systems for example, hexane, t-butylmethylether, acetonitrile, water (H*ter*AcWat) (Friesen, et al., 2007). The overwhelming majority of systems contain water. Whether this aqueous phase is the upper or the lower phase is a function of the density of the solvents used. In most solvent systems, the aqueous phase will be the lower phase. However, in the case of chlorinated solvent systems, for example, water and dichloromethane (DCM), the water will be the upper phase and the DCM the lower phase due to their density. All these examples of solvent systems contain organic solvents and water so they will not be appropriate for separating certain types of molecules, for example, proteins and other biological molecules, which would be denatured by the organic solvent. To overcome this problem, aqueous two phase systems (ATPS) have been developed. These can consist of either two polymer solutions or a polymer/salt solution (Zaslavasky, 1995). A common example is PEG1000 and potassium phosphate. These biphasic systems are more environmentally friendly as they do not contain volatile organic solvents. There are also examples in the literature of non-aqueous systems such as heptane and dimethyl sulfoxide (DMSO) (Berthod, et al., 1996) and heptane and acetonitrile (Conway).

Like compounds with high K_d values, ionisable analytes can be difficult to separate using traditional CCC/CPC solvent systems. However, by using pH modifiers, extracting reagents or salting-out (in) agents added to the solvent system, the successful separation of these ionisable compounds has been achieved. One such technique is pH-zone-refining. This involves the addition of acid to the mobile phase and the addition of base to the stationary phase resulting in pH gradient in the column and the elution of very sharp, well defined peaks. This is advantageous as the fractions are highly concentrated and the solvent system can tolerate up to ten times as much sample loading as a regular solvent system (Friesen, et al., 2007). The three main systems used with the addition of acid and base are t-butyl methyl ether (MTBE)/water, MTBE/acetonitrile/water (4:1:5 or 2:2:3) and 1-butanol/water (Ito, 2013).

1.5.1. The HEMWat system

The most widely used solvent systems are the HEMWat family which are made up of four solvents: heptane, ethyl acetate, methanol and water. By changing the ratios of the solvents, this solvent system can be used to separate out compounds with a large polarity range. Changing the proportions of the each solvent will change the polarity of the overall system and the polarity difference between phases. Polarity is the ability of a substance to be attracted or repelled by electrostatic forces and this control over the polarity allows the system to be adjusted to optimise the partitioning of many different compounds. The HEMWat systems were originally proposed by Oka (Oka, et al., 1991) and was further developed into the Arizona table (Margraff, et al., 1994) and the HEMWat table (Garrard, 2005). The low boiling points of the four solvents mean that the solvents used can be removed easily leading to the faster recovery of the compounds being separated. Garrard et al. demonstrated that the recovered solvent system could be recycled and used again by adjusting phase compositions to make up the original HEMWat system with the minimal addition of solvents (Garrard, et al., 2007). In addition, their low viscosity makes them good candidates for a solvent system in CCC/CPC. Heptane, methanol and water show no absorbance above 205 nm and although ethyl acetate can have absorbance up to 254 nm, the absorbance is usually small enough to allow the use of a UV detection method to monitor the separation (Lu, et al., 2009). HEMWat also has the added flexibility that heptane can be exchanged for hexane or isooctane without adversely affecting the properties of the solvent system (Berthod, et al., 2005).

Garrard developed a systematic screening method for the HEMWat solvent systems (Garrard, 2005) using a liquid handling robot and a numeric labelling scale from 7 - 28 to denote polarity, within which HEMWat 7 was the most polar and HEMWat 28 was the least polar. The composition of each of the HEMWat systems can be found in Table 1. The compositions of each phase of the HEMWat systems were determined using Gas Chromatography (GC) (Garrard, 2005). The HEMWat systems denoted 1-6 contain butanol and not always the other four HEMWat solvents. The upper phase contains very little water and methanol across the entire series and the percentage composition of the ethyl acetate decreases from HEMWat 6 to HEMWat 28 at the same rate as the percentage composition of heptane increases. The lower phase of the HEMWat systems contain very little heptane until HEMWat 25 and the amount of

ethyl acetate in the lower phases is relatively stable, peaking between systems 15 and 20 (Figure 4 and Figure 5) (Garrard, 2005).

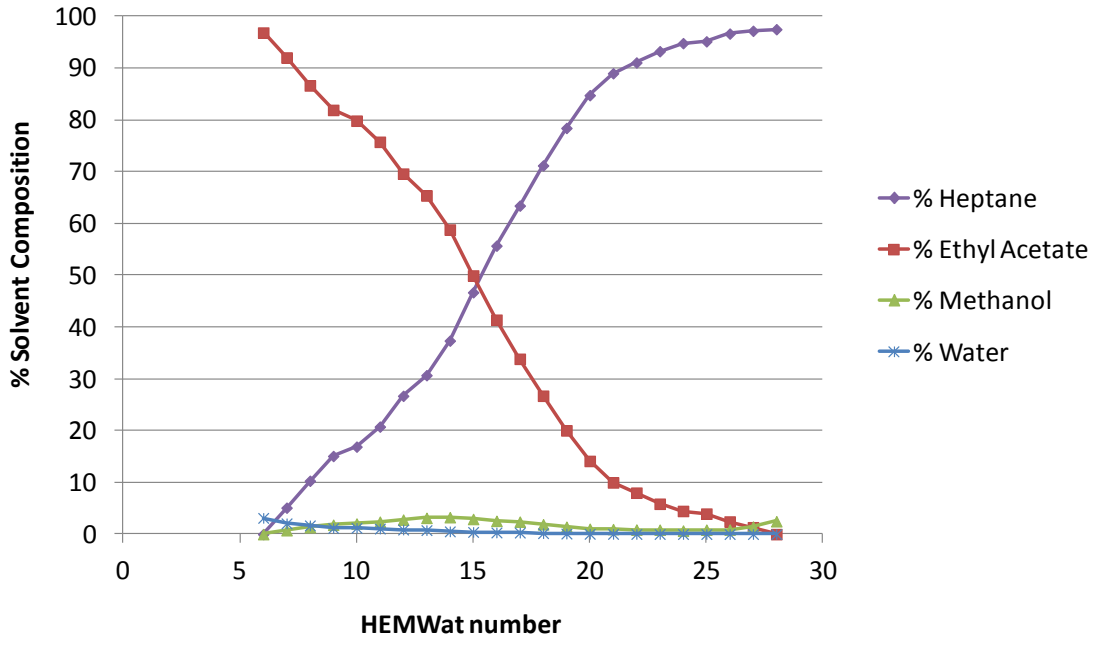


Figure 4 - The percentage composition of upper phase of the HEMWat systems across the polarity range of all 28 systems.

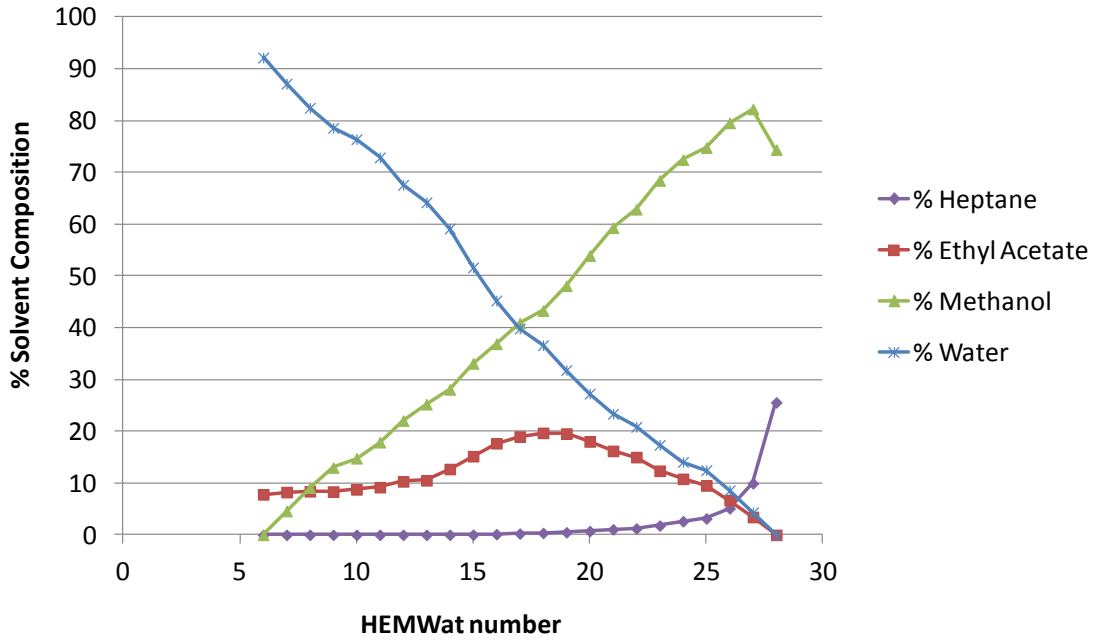


Figure 5 – The percentage composition of lower phase of the HEMWat systems across the polarity range of all 28 systems.

Table 1 - The volume ratios of the five solvents required to make up each HEMWat system. HEMWat systems 1-6 are not HEMWat systems but can be used to extend the polarity range. The polarity of the HEMWat systems increases towards HEMWat 7 and decreases towards HEMWat 28 (Garrard, 2005).

No	Heptane	Ethyl Acetate	Methanol	Butanol	Water
1	0	0	0	5	5
2	0	1	0	4	5
3	0	2	0	3	5
4	0	3	0	2	5
5	0	4	0	1	5
6	0	1	0	0	1
7	1	19	1	0	19
8	1	9	1	0	9
9	1	6	1	0	6
10	1	5	1	0	5
11	1	4	1	0	4
12	1	3	1	0	3
13	2	5	2	0	5
14	1	2	1	0	2
15	2	3	2	0	3
16	5	6	5	0	6
17	1	1	1	0	1
18	6	5	6	0	5
19	3	2	3	0	2
20	2	1	2	0	1
21	5	2	5	0	2
22	3	1	3	0	1
23	4	1	4	0	1
24	5	1	5	0	1
25	6	1	6	0	1
26	9	1	9	0	1
27	19	1	19	0	1
28	1	0	1	0	0

One of the important things about the HEMWat systems is the relatively uniform polarity gap between each of the numbered systems, as well as between the two phases of each system. Polarity is important in partitioning and there are many different scales and ways to quantify it.

1.6. Polarity

Polarity is the force caused by an uneven spread of electrons around a molecule. There have been many attempts to quantify this and many different polarity scales have been proposed, with Katritzky et al. listing 184 (Katritzky, et al., 2004). These scales can be based on experimental measurements, for example equilibrium, kinetic or spectroscopic measurements or multiparameter approaches. However, the polarity scales are mainly aimed at pure solvents, which limit their accuracy with Reichardt stating that there was no parameter capable of entirely describing solvent polarity (Reichardt, 1979).

1.6.1. Reichardt polarity

Reichardt developed a polarity scale based on spectroscopic measurements. The Reichardt polarity parameter (E_T) is defined as "the molar electronic transition energies (E_T) of dissolved, negatively charged solvatochromic pyridinium N-phenolate betaine dye (Figure 6), measured in kilocalories per mole (kcal/mol) at room temperature (25°C) and normal pressure (1 bar)" (Reichardt, 1994). This can be numerically demonstrated using Equation 8.

$$E_T = hcv_{max}N_A = 2.8591 \times 10^{-3} = \frac{28591}{\lambda_{max}}$$

*Equation 8 - Reichardt Polarity where h is the Planck's constant (6.626 x 10-34 m^2 kg s^{-1}), c is the speed of light (2.99 x 10^8 m s^{-1}), v_{max} is the frequency of the maximum of the longest wavelength, intramolecular charge-transfer π-π*absorption band of dye, N_A is the Avogadro's number (6.02 ×10^{23} mol^{-1}), λ_{max} is the wavelength of the maximum of the longest wavelength.*

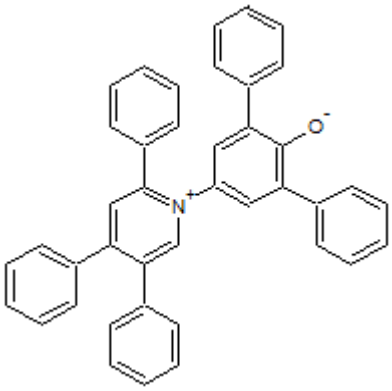


Figure 6 – Molecular structure of Reichardt dye (pyridinium N-phenolate betaine) (Reichardt, 1994).

The polarity of a solvent is determined by dissolving the dye in the solvent and measuring its absorbance in an Ultraviolet-Visible (UV-Vis) spectrometer. The E_T value can be normalised (E_T^N) using water as a polar solvent and the tetramethylsilane (TMS) as a nonpolar solvent resulting in a normalised scale from 0 at TMS to 1 at water (Equation 9).

$$E_T^N = \frac{E_T(solvent) - E_T(TMS)}{E_T(water) - E_T(TMS)}$$

$$E_T^N = \frac{E_T(solvent) - 30.7}{32.4}$$

Equation 9 - Normalised Reichardt Polarity where E_T(solvent) is the molar electronic transition energy of solvent, E_T(TMS) is the molar electronic transition energy of solvent of tetramethylsilane and E_T(water) is the molar electronic transition energy of water.

The pyridinium N-phenolate betaine dye (Figure 6) is not always soluble in nonpolar solvents. However, the linear correlation between E_T and λ_{max} can be exploited using Equation 10 for solvents with low dye solubility.

$$E_T = \frac{\frac{28591}{\lambda_{max}} - 1.808}{0.9424}$$

Equation 10 – Reichardt equation (Reichardt, 1965) where E_T is the Reichardt polarity and λ_{max} is the maximum wavelength.

This method has been successfully used to measure the polarity of some solvent systems in CCC/CPC (Abbott, et al., 1991). However, in certain systems, for example HEMWat, the dye's lack of solubility in hydrocarbon and water become inhibitory at

the extreme ends of the polarity scales. The dye can be altered to be more water soluble (Figure 7).

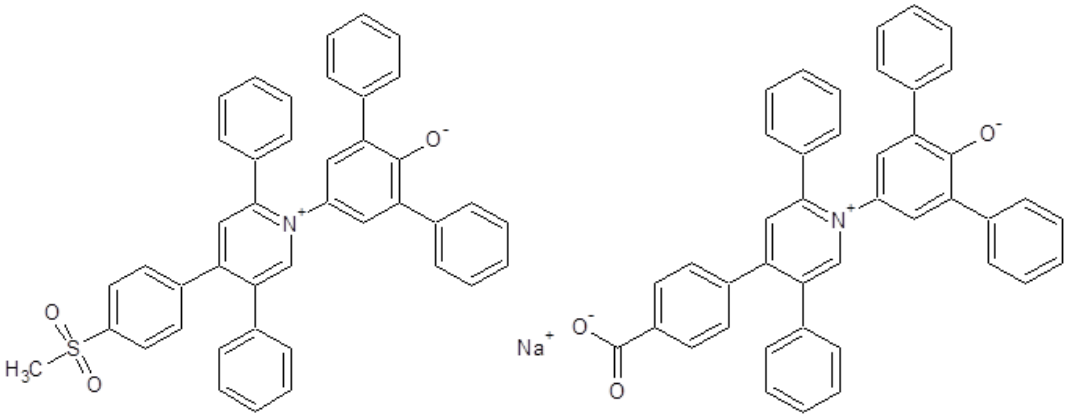


Figure 7 – Two examples of the structure of Reichardt dyes with enhance solubility in polar solvents

1.6.2. Hildebrand polarity

The Hildebrand solubility parameter (δ) is the energy required to make a cavity in the solvent (Kamlet, et al., 1981). It is defined in Equation 11.

$$\delta = \left(\frac{\Delta_{vap}E_m}{V_m}\right)^{\frac{1}{2}}$$

Equation 11 - Hildebrand polarity (δ) where δ is the Hildebrand polarity, $\Delta_{vap}E_m$ is the molecular energy of vaporisation and V_m is the molar volume.

The heat of vaporisation and the molar volume of the liquid phase are the only parameters needed to calculate the Hildebrand solubility parameter. The heat of vaporisation can be measured through calorimetry, whilst the molar volume can be obtained through pressure measurements. This polarity scale has some disadvantages as the Hildebrand solubility parameter of a solid compound is hard to obtain. A second disadvantage to this scale is that the assumption of solution ideality leads to poor predictions of solubility in polar solvents (Martin, et al., 1980).

1.6.3. Snyder polarity

Snyder defined polarity as the “ability of a sample molecule or solvent molecule to interact”. This scale is based this on four interactions between solute and solvent: dipole, dispersion, dielectric and hydrogen bonding (Snyder, 1968). An equation was developed using the model of solvent molecules adsorbing onto a surface (Equation

12) with the conditions that there is a low water content and the molecule is being absorbed onto highly activated silica.

$$\varepsilon_0 = -\frac{\Delta G_0}{2.3RTA_M}$$

Equation 12 – The Snyder polarity (ε_0) where ε_0 is the Snyder polarity (ε_0), ΔG_0 is the variation in the free adsorption energy of 1 mole of mobile phase, R is the molar gas constant (8.314 m^2 kg s^{-2} K^{-1} mol^{-1}), T is the Temperature (K) and A_M is the area of adsorbent occupied by 1 mole of mobile phase.

1.6.4. Rohrschneider and Snyder polarity

Rohrschneider measured the gas-liquid partition coefficients of six solutes in 81 solvents with a wide range of polarities. The six solutes were: n-octane, toluene, ethanol, dioxane, nitromethane, and butanone. Snyder used these gas-liquid coefficients measured by Rohrschneider to produce polarity parameters. This scale became known as the Rohrschneider and Snyder polarity scale and is based on the summation of the polarity of the three interaction parameters; hydrogen bonding ability, hydrogen donating ability and the dipole-dipole interaction (Poppe, et al., 1981).

1.6.5. Kamlett-Taft

The Kamlett-Taft polarity (π_S^*) scale is based on solvent induced shifts of the longest wavelength absorption for seven nitro aromatic indicators. It was initially based on 29 solvents but was refined by Buncel using UV/Vis spectral data to normalise the scale so that π_S^* for cyclohexane is 0.00 and 1.00 for DMSO (Reichardt, 1994).

$$\pi_S^* = \frac{\left(\nu_S - \nu_{c-C_6H_{12}}\right)}{\left(\nu_{DMSO} - \nu_{c-C_6H_{12}}\right)}$$

Equation 13 - Kamlet-Taft equation where π_S^* is the Kamlet-Taft polarity of a compound in solvent, S, ν_s is the maximum frequency of the solvatochromic absorption band of 4-methylnitrobenzene in solvent, S, $\nu_{c-C_6H_{12}}$ is the maximum frequency of solvatochromic absorption band in cyclohexane and ν_{DMSO} is the maximum frequency of solvatochromic absorption band in DMSO.

It is divided into three parts: α which is the hydrogen bond donor term, β which is the hydrogen bond acceptor term and π which is the dipolarity/ polarizability term (Kamlet, et al., 1981).

1.6.6. Dipole moment, p

The dipole moment, p (Mills, et al., 1993) results from the non-uniform distribution of electrons around a molecule (Equation 14).

$$p = \frac{T}{E}$$

Equation 14 - Dipole Moment polarity where p is the electric dipole moment, T is the torque and E is the electric field strength.

Torque is the sum of moments of forces not acting along the same line. The larger the difference in charge over the molecule, the larger the dipole moment. Molecules with large dipole moments are described as electronegative, with elements in the upper right hand corner of the periodic table likely to increase a molecule's electronegativity.

1.6.7. Relative Permittivity, ε_r

Relative permittivity, ε_r, (formally known as the dielectric constant) is the "amount of electrical energy stored in a material by an applied voltage, relative to that stored in a vacuum" (Mills, et al., 1993). It provides an indication of whether a material can become polarised under an electric field. The higher the value of relative permittivity, the higher the polarity of a substance. This measure is temperature, pressure and frequency dependent, with all molecules having electronic and atomic polarisation and polar molecules having additional orientational polarisation. The permittivity of free space is $\varepsilon_0 = 8.85 \times 10^{-12}$ m^{-3} kg^{-1} s^4 A^2 which is taken as a fixed constant.

1.7. Prediction of Partition Coefficients

The choice of solvent system is normally based on an analyst's past experience, trial and error and/or literature analysis. This may mean that systems that would give very well defined chromatography are missed or that large quantities of time and solvent are used selecting an appropriate solvent system. Being able to predict the K_d values of compounds would allow prediction of the optimised CCC solvent system without time consuming, solvent intensive experiments to experimentally determine the K_d value of compounds in many solvent systems.

1.7.1. Experimental prediction of Partition Coefficients

There are many practical ways of choosing a solvent system. One method for screening solvents is to run a trial separation on a small column so that K_d values can

be calculated from the CCC chromatogram to give the analyst a feel for the polarity of the compound (Hu, et al., 2012). However, this is time consuming and if the compound is expensive or only available in low amounts, this is a disadvantage.

An alternative method for experimentally determining K_d values was put forward by Friesen and Pauli in 2007 called the “GUESSmix” method (Friesen, et al., 2005). This method involved carrying out Thin Layer Chromatography (TLC) on 22 reference compounds in many different solvent systems. The retention factor (R_f) of the structurally diverse reference compounds can be compared to the R_f value of the compound to be separated so that a compatible solvent system can be chosen. However, this method requires TLC analysis for the compound and this is also time consuming and requires compound which may be expensive or scarce.

Foucault and Chevolot developed the “best solvent method” (Chevolot, et al., 1988) based on initial identification of a solvent that could dissolve a large amount of compound to be separated. A two phase system in which this solvent partitioned between the two phases was then identified, using ternary diagrams from Sørensen and Arlt (Sørensen, et al., 1980). This was the system used to perform the separation. This method requires a large amount of compound to test its solubility, which may be difficult if the compound is valuable or scarce.

Han et al. (Han, et al., 2008) developed a method for partition coefficient prediction using the solvent system n-hexane/methanol/water. Initially, the ratio of n-hexane and water was kept constant and the amount of methanol was changed. K_d was plotted against the volume of methanol for four different n-hexane/water systems. Two trend lines were added; one with an exponential relationship and one with a power relationship. These lines were used to predict K_d values. The experiments were repeated with a system that had the total volume of methanol and water fixed as double the volume of n-hexane. The experimental K_d values were obtained and compared to the predicted values from the two graphs. It was found that the exponential trend line predicted K_d values that were closer to the experimental values. The predictions were also made for HEMWat with the total volume of n-hexane and ethyl acetate kept equal to the total volume of methanol and water. This time the power equation for the plot of K_d against the volume of methanol performed the best when predicting K_d values. However, there were some examples of the power equation

performing very poorly and the linear line giving the best relationship. This method will be very time consuming as many experimental measures are needed to assess which trend line is the best. The other disadvantage of this method is that there is no guidance for whether a linear, exponential or power equation will allow the most accurate predictions (Han, et al., 2008).

Dubant et al. developed a solvent screening method using design of experiments (DoE) to limit the experimental work that must be carried out. Nine partitions were measured across a range of solvent system compositions. The polarity of the solvent systems was used to produce a 3D polarity map which was combined with statistical software, to predict the composition that will result in the compound having a partition coefficient of one (Dubant, et al., 2008). This method is time consuming as it involves both experimental and computational work for the end user to predict the optimal solvent system.

The capillary wavelength of a solvent system can also be used to determine its suitability for use in CCC. It is caused by disturbances along the boundary between the two phases of a solvent system. Solvent systems with a short capillary wavelength have a higher stationary phase retention which is advantageous when using CCC to perform a separation (Fedotov, et al., 2000). However, the capillary wavelength cannot be measured directly so must be calculated using Equation 15 (Menet, et al., 1994). This equation requires the interfacial tension between the two immiscible phases (γ) and their density difference (Δρ). The method used in the paper for measuring interfacial tension involves two, six hour steps. This is not practical for swift solvent selection.

$$\lambda_{cap} = 2\pi \sqrt{\frac{\gamma}{|\Delta\rho| g}}$$

Equation 15 - λ_{cap} is the capillary wavelength, γ is the interfacial tension between the phases, Δρ is the density difference between the phases and g is gravitational force (9.81 m/s^2).

An alternative screening method was developed which involved measuring the partition coefficients of compounds in three broadly spread HEMWat solvent systems. This was used to narrow the area where it was likely that the compound to be separated would have a reasonable K_d value. Once the solvent system in which the

compound has the most acceptable K_d value was identified, three further solvent systems near to this identified solvent system were then screened. If this produced a solvent system in which the compound has a K_d value of between 0.5 and 2 then this was used to run the CCC separation. If needed, a further three solvent systems could be screened guided by the results of the previous K_d measurements (Lu, et al., 2009).

1.7.2. Computational Prediction of Partition Coefficients

A computer model that requires only the molecular structure of a compound would be a much less labour intensive solution to predicting K_d values allowing a novice to carry out a separation using CCC/CPC. There are currently several computational methods for predicting partition coefficients; however, they have certain limitations.

Hopmann et al. used the software COSMO-RS (COnductor-like Screening MOdel for Real Solvents) to calculate the activity coefficients of the upper and lower phases respectively. This could then be used to predict the partition coefficient which in turn could be used to identify the optimal solvent system for separation. The disadvantage to this method is that the conformation of the molecule plays a very important role in the calculation. Therefore, the conformation must be exactly calculated which is computationally expensive, especially for larger molecules, such as some natural products (Hopmann, et al., 2011). If the molecular structure of a compound is not known, this method would not be applicable (Ren, et al., 2013). Hopmann et al. then subsequently published another paper in which they used COSMO-RS to predict the partition coefficient in CCC solvent systems (Hopmann, et al., 2012). They focused on the HEMWat system which contains heptane, ethyl acetate, methanol and water. The assumption was made that industry prefer a small list of solvents, so ten solvents were selected and modelled. This is a potential disadvantage as the best solvent system may not have been investigated. The method is similar to the “best solvent approach” (Chevolot, et al., 1988) and involves predicting the solubility of a compound. The solvent system was selected based on this “best solvent’s” partitioning between the two phases. The partition coefficient for a range of compositions of that system was then calculated and used to predict the best solvent system. However, Hopmann et al. state that COSMO-RS is only able to predict solubility for neutral compounds in non-aqueous systems. This limits the applicability of the approach since most of the solvent systems in CCC/CPC contain water. Furthermore, experimental solubility measurements require lots of compound which is often not available.

An alternative to COSMO-RS is UNIFAC (Universal quasichemical functional-group activity coefficients) software (Li, et al., 2003). This programme uses thermodynamics to calculate K_d as the entropic component of ΔG is an important factor in the positioning of equilibrium i.e. K_d. The big disadvantage of this programme is that solvents that are partially soluble in each other (like methanol and water) affect the thermodynamic nature of the system, which UNIFAC cannot take into account. The UNIFAC software also fails to allow for the additional change in the partition coefficient that electron withdrawing groups produce (Leo, et al., 1971). A potential disadvantage of this programme is its use of group interaction parameters which are not always available. Ren et al. used NRTL-SAC (Non-Random Two Liquid – Segment Activity Coefficient), UNIFAC and GA (Generic Algorithm) to predict partition coefficients for solvent system selection in CCC (Ren, et al., 2013). The predictions from NRTL-SAC are based on 4 descriptors: hydrophobic, hydrophilic, attractive and repulsive. UNIFAC is used to generate the composition of solvent systems. The activity coefficient was calculated using regression, which was combined with the results from UNIFAC using GA in MATLAB (matrix laboratory) software. Experimental K_d values from five different compositions were then used to predict the K_d values for other compositions of the solvent system. The composition with the predicted K_d value nearest to one is then chosen to run on the CCC. This method is not purely computational as some experimental K_d values are needed for the prediction. This is a disadvantage if the compound to be separated is expensive or supply is limited.

The literature also contains more specific examples of modelling partition coefficients. The partition of a compound between octanol and water (logP) is used in the pharmaceutical industry to model the blood brain barrier (Levin, 1980). LogP is also a direct measure of polarity (hydrophobicity) and is very sensitive to polarity changes over a large range (Laane, et al., 1987). The logP of a mixture of two compounds can be described as Equation 16.

$$logP_{mixture} = X_1 logP_1 + X_2 logP_2$$

Equation 16 – Octanol/water partition coefficient of a mixture where X_1 and X_2 are the mole fractions of each compound and logP is the octanol/water partition coefficient.

Amat et al. proposed a method predicting logP (octanol/water) using Molecular Quantum Similarity Measures (MQSM). It is based on comparing the electron density

of the compound to be separated with the electron density of a compound with known logP values (Amat, et al., 1998). LogP has also been predicted using neural networks (NN) (Livingstone, et al., 2001) and Scalable Process Architecture (SPARC) (Hilal, et al., 2004). SPARC uses a combination of SARs (Structure Activity Relationships), Linear Free Energy Relationships (LFER) and PMO (Perturbational Molecular Orbital) theory. SPARC is capable of calculating its own molecular descriptors but can use parameters that have been generated using a different model. It is a fragment based method and is able to distinguish between sterically different molecules. The method calculates partition coefficients using activity coefficients (γ) which are calculated using a Flory-Huggins term and Gibbs free energy, and the ratio of the molecularities of the two phases (Equation 17).

$$-RT\log\gamma_{ij}^{\infty} = \sum_{Interaction}^{All} \Delta G_{ij}(Interaction) + RT\left(-log\frac{V_i}{V_j} + \frac{\left(\frac{V_i}{V_j}-1\right)}{2.303}\right)$$

Equation 17 - R is the molar gas constant (8.314m² kg s⁻² K⁻¹ mol⁻¹), T is the temperature (Kelvin), γ is the activity coefficient, ΔG is the Gibbs free energy, V_i is the cavity volume of the solute and V_j is the cavity volume of the solvent.

However, this method does not perform well when predicting K_d values for large hydrophobic molecules and, like the UNIFAC method, cannot take into account the additional effect, which electron withdrawing groups have on the partition coefficient (Leo, et al., 1971). The disadvantage of this method is that many systems are not completely immiscible, which is an essential criterion for the use of this method.

1.8. Quantitative Structure Activity Relationship models

Quantitative Structure Activity Relationship (QSAR) models use molecular descriptors to predict biological activity data (Kubinyi, 1997). The larger the number of molecular descriptors, the higher the likelihood of the QSAR model being able to successfully describe the relationship between the molecular structure and the biological activity. There are 1600 properties available in the molecular descriptor library for 2D QSAR calculations (Du, et al.). Some descriptors are molecular and atom based such as the number of oxygen atoms present, others are from quantum chemical calculations such as the largest maximum eigenvalue from a connectivity matrix or measured from spectroscopy such as Abraham's parameters. Descriptors that are important in the

relationship between the compounds' molecular structure and the property being modelled are deemed "significant". The significant descriptors are identified from patterns within a database of known experimental data. The ability to generate large consistent experimental data sets, through high-through-put data generation avoids the problem of combining data from many sources. This is important as the resulting QSAR model is only as accurate as the data used to train it. The use of large training sets of data increases the size of the experimental error that will be tolerated by the model. However, large training sets are not always necessary as successful QSAR models have been built with a training set as small as 44 compounds (Patil, et al., 2013). Experimental data obtained from an HPLC have been used in conjunction with the QSAR methodology to build a wide variety of models including antitumour (Koba, et al., 2012), antifungal (Niewiadomy, et al., 1998), antimicrobial, antimycotic and tuberculostatic (Jozwiak, et al., 1999). QSAR modelling is also a widely used tool in computer aided drug design due to its ability to predict the impact of a small structural change on biological activity (Du, et al.).

1.8.1. Historical Development of the QSAR Methodology

As early as 1868 while investigating substances poisonous to humans, Crum-Brown and Fraser (Brown, et al., 1868) identified that some physiological actions were related to chemical composition, noting that methyl amine (CH_3NH_2) is inert whilst hydrogen cyanide (HCN) is highly poisonous, despite both being made up of the same three elements. This was further extended by Meyer (Meyer, 1899), who linked the narcotic action of compounds to their olive oil/water partition coefficients. In 1938, Hammett noticed a linear relationship between the equilibrium behaviour of ester hydrolysis using different alcohols with the equilibrium behaviour of benzoic acid (Hammett, 1938). This relationship connected equilibrium constants with chemical structure for the first time with Hammett deriving Equation 18.

$$\rho\sigma = \log\frac{K}{K_0}$$

Equation 18 – The Hammett Equation where ρ is the reaction constant, σ is the substituent constant, K is the equilibrium constant for substituted reactants and K_0 is the equilibrium constant for non-substituted reactants.

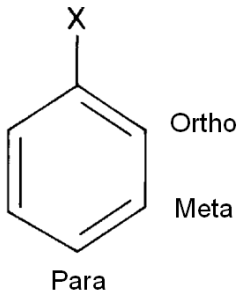


Figure 8 - Benzene with substituent X on carbon 1. Carbon 2 is known as ortho carbon, carbon 3 is known as the meta carbon and carbon 4 is known as the para carbon

Hammett found that by altering the nature of a substituent X on a molecule (Figure 8), the position the next substituent is added to, can be controlled. Substituent X can be either electron rich or electron poor. Electron rich substituents (e.g. -OH) will encourage addition on the ortho- and para- carbons. Electron poor substituents (e.g. $–NO_2$) will encourage addition on the meta- carbons. There are two mechanisms which govern this: the inductive effect and resonance effects.

The inductive effect is the donation or withdrawing of electron density from the carbon atom of a carbocation intermediate by the adjacent functional group. Electron donating groups (EDG) donate electrons to the carbocation intermediate during a reaction stabilising the intermediate which encourages the reaction. An example of this is shown in Figure 9.

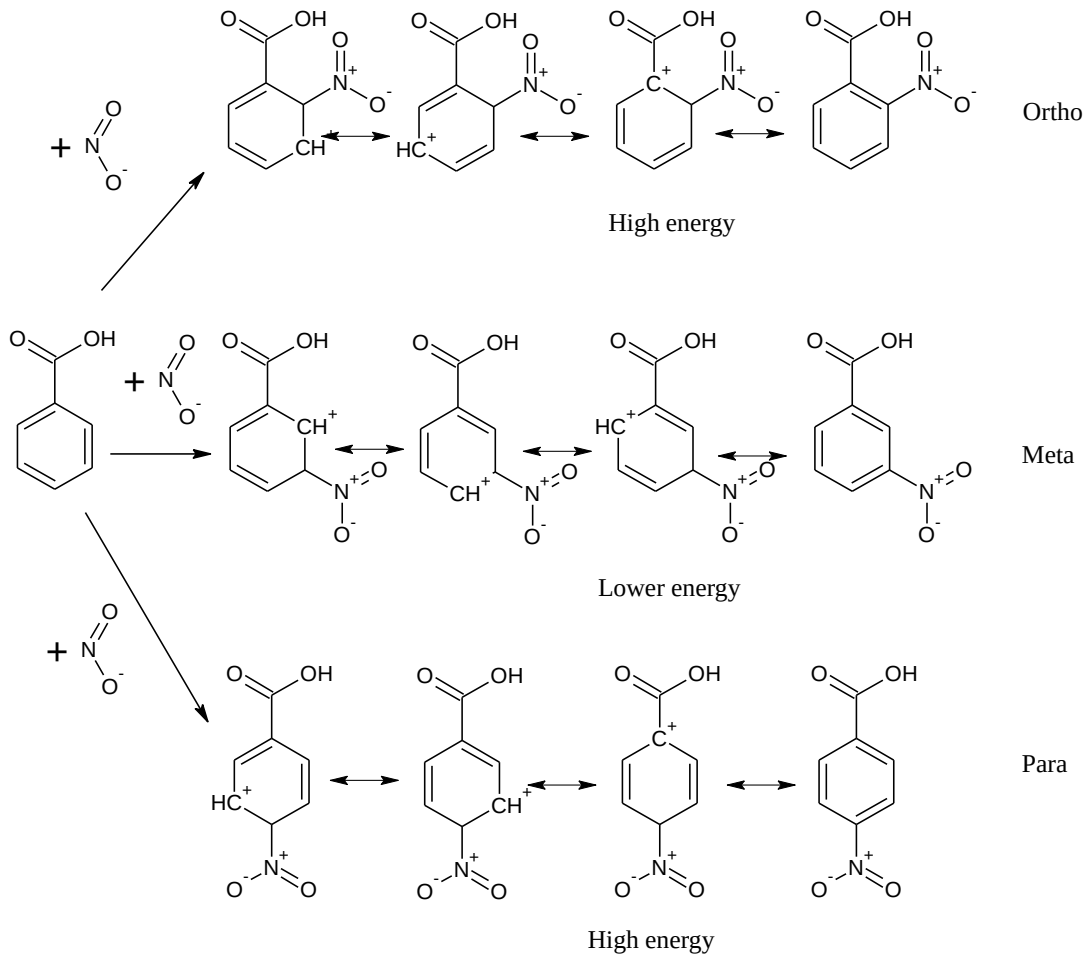


Figure 9 – The carboxylic aid group is meta directing. The meta- position avoids a positive charge on the carbon of the carboxylic acid group which is a high energy molecule. Therefore, the majority of the product will be the meta form.

The substituent constant, σ, value in the Hammett equation (Equation 18) is always negative for an EDG substituent. Electron withdrawing groups (EWG) remove electron density from the carbocation intermediate, destabilising the intermediate discouraging the reaction. The σ value in the Hammett equation (Equation 18) is always positive for a EWG substituent.

Taft identified that the Hammett equation failed when applied to aliphatic systems and ortho- substituted benzene in 1952. Steric effects were included to derive the Taft equation (Equation 19) by adding the Taft steric parameter to the Hammett equation (Taft, 1952), which is always less than or equal to 0. The polar, steric and resonance effects were separated as shown in equation below.

$$\log K = \log K_0 + \rho^* \sigma^* + \delta E_S$$

Equation 19 – The Taft Equation where ρ is the susceptibility of a reaction to the electronic nature of substituents, σ* is the polar substituent constant, δ is the measures sensitivity of the studied reaction to steric effects of substituents, E_s is Taft's steric parameter.*

In 1962, Hansch adapted the Hammett equation for use in biological systems (Hansch, et al., 1962) developing a structure-activity relationship between plant growth regulators, Hammet's constant and hydrophobicity. This led to Equation 20 which relates the physicochemical properties of molecules correlated with their biological activity.

$$\log\left(\frac{1}{C}\right) = a\pi + b\sigma + cE_s + k$$

Equation 20 – The Hansch equation where C is the molar concentration, π is the lipophilicity parameter, σ is the electronic parameter, E_s is Taft's steric parameter and a, b, c and k are constants.

Hansch found that not all relationships were linear so introduced a non-linear version of the equation (Equation 21). This equation allows the description of the relationship dependent on a linear and non-linear property e.g. drug transport is linearly dependent on binding affinity whilst being non-linearly dependent on lipophilicity.

$$\log\left(\frac{1}{C}\right) = a\pi + b\pi^2 + c\sigma + dE_S + k$$

Equation 21 – The non-linear Hansch equation where C is the molar concentration, π is the lipophilicity parameter, σ is the electronic parameter, E_s is Taft's steric parameter and a, b, c and k are constants (Kubinyi, 1997).

In the 1960s, Free and Wilson produced an alternative model which also used structural features to predict biological properties (Free, et al., 1964). The Free Wilson model uses the sum of the activity of individual substituents and the biological activity of a reference compound, to calculate biological activity. For example, this method was used to model the ability of N,N-dimethyl-a-bromophenethylamine and its analogues, to inhibit the effect of epinephrine and norepinephrine.

$$\log\left(\frac{1}{C}\right) = \sum\sum GX + \mu$$

Equation 22 - Free Wilson Model where C is the molar concentration, G is the group contribution, X is the structural feature and μ is the biological activity of reference compound (usually the parent molecule).

The advantage of the Free Wilson method is that it requires only the biological activity and molecular structure to make predictions. However, the biological activity for a minimum of two parent molecules with at least two substituents in two different positions, are needed to make the predictions.

Free Wilson is used less frequently than the Hansch equation as it can only be used in its linear form and is prone to overfitting (section 1.8.2.2) (Du, et al.). The Hansch equations can be used in a parabolic form which gives this equation added flexibility (Kubinyi, 1988). However, if both of the methods are in linear form, they can be combined and compared. They must first be normalised using the Fujita-Ban Model which relates all activity contributions to hydrogen using Equation 23 (Fujita, et al., 1971).

$$\log\frac{1}{C} = \sum k\Phi + c$$

Equation 23 – The combination of Free-Wilson and Hansch models using the Fujita-Ban normalisation where C is the molar concentration, k_j is the coefficient, Φ_j are the physico-chemical parameters and c is a constant.

1.8.2. Descriptors

To determine relationships between molecular structure and biological activity, a way must be found to describe the molecule. This can be done using a range of descriptors, for example, number of atoms, size, and charge.

1.8.2.1. Abraham Parameters

A set of parameters were explored by Abraham (Taft, et al., 1985) as a possible method for the prediction of $\log K_d$ values. The five parameters were chosen as a diverse selection of descriptors to increase the probability of the equation being able to successfully model a property. Two of the parameters were based on hydrogen bonding. Hydrogen bonding acidity (A) is a measure of the hydrogen bond donor ability of the compound, and hydrogen bonding basicity (B) is a scale by which the willingness

of a compound to be a hydrogen bond acceptor is measured. The third parameter is polarity/polarisability (S), which is the force caused by an uneven spread of electrons around a molecule. Another parameter is based on electrons; excessive molar refraction (E) is the indication of the solute-solvent interaction that arises through the presence of polarisable electrons in a molecule. The final parameter is McGowan volume (V), a measure of the cavity effect, which is the energy required to disrupt solvent-solvent bonds. These parameters can be experimentally determined or predicted. Parameters E and V are determined from the structure of a molecule whilst A, B and S are experimentally determined. Parameter E is calculated from a molecule's refractive index and parameter V is from atom bond contributions. Parameters A and B are obtained from water-solvent partitions, whilst parameter S is determined using gas or liquid chromatography. All the parameters apart from E can be predicted to a high level of confidence using the COSMO-RS software, minimising the need for experimental measurements (Jover, et al., 2004). All five parameters can be obtained using the National Chemical Database ILab website which will provide predicted values using the Absolv software (Advanced Chemistry Development Labs, Toronto, Ontario, Canada) if the experimental values have not been determined.

Abraham found (Abraham, et al., 2004) that by using regression to produce coefficients for the summation of the five parameters, predictive models could be created for a large range of molecular properties, e.g. toxicity (Equation 24).

$$SP = c + \alpha A + \beta B + sS + vV + eE$$

Equation 24 – Abraham's equation where SP is some property, A is the hydrogen bonding acidity, α is the coefficient for hydrogen bonding acidity term, B is the hydrogen bonding basicity, β is the coefficient for hydrogen bonding basicity term, S is the polarity/ polarizability, s is the coefficient for polarisability term, E is the excessive molar refraction, e is the coefficient for molar refractivity term, V is the McGowan volume, v is the coefficient for McGowan volume term and c is the constant.

The coefficients in Equation 24 are used to characterise phase and contain chemical information (Zissimos, et al.). The α coefficient is the hydrogen bond acidity and the β coefficient if the hydrogen bond basicity. The e coefficient is the "tendency of phase to interact with solutes through π and n electron pairs". It is likely to be a positive value; however fluorine can cause it to be negative due to its extreme electronegativity. The s coefficient is the "tendency of the phase to interact with dipolar/polarisable solutes".

The v coefficient is the “measure of hydrophobicity of phase”. It is used to describe the dispersion interactions and cavity forces that are important in solubility (Jover, et al., 2004).

1.8.2.2. AZ Molecular Descriptors

C-Lab (Internal AstraZeneca software, Alderley Park, Cheshire, UK) can generate 196 2D molecular descriptors, these can be divided into seven main categories: lipophilicity, hydrogen bonding, size and shape, charge and polarity, atom counts, topology and druggability. The druggability of a molecule is its ability to modulate a target in vivo. The parameters are generated using SMILES (Simplified Molecular-Input Line-Entry System) (Weininger, 1988) so are purely based on molecular structure. However, using all 196 descriptors has a tendency to lead to overfitting of the model. Overfitting describes a model that is well fitted to the training set data but has no predictive capability. QSAR models that are overfitted include descriptors that are of no value, sometimes at the expense of descriptors that are valid. To avoid this, the “Top 14 AZ 2D molecular descriptors” are mainly used (section 9.2).

1.8.3. Mathematical methods for building QSAR models

Regression is the most common mathematical method for QSAR generation (Selassie, et al., 2010). Four common methods for building QSAR models include multiple linear regression (MLR) and partial least squares (PLS), which are both multivariate analysis methods and support vector machines (SVM) and random forest (RF) which are machine learning techniques. The regression methods are not limited to these four techniques, for example neural networks (NN) have been successfully used to produce QSAR models (Yan, et al., 2010) for the melting points for imidazolium bromides and imidazolium chlorides ionic liquids.

1.8.3.1. Multivariate Analysis Methods (MVA)

Multivariate Analysis (MVA) allows the analysis of more than one statistic at the same time. The two methods investigated were Multiple Linear Regression (MLR) and Partial Least Squares (PLS). These methods model the relationship between multiple independent variables, X and the dependent variable, Y. Both models make the assumption that both X and Y can be modelled using the same variables.

1.8.3.1.1. Multiple Linear Regression (MLR)

Multiple Linear Regression is used to identify multiple descriptive variables that can be used to describe a dependent variable. In a data set with n data points, the multiple linear regression can be denoted as Equation 25.

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_n x_n + \varepsilon_i$$

Equation 25 – Regression equation where y is the dependent variable, x is the independent variables, the β terms are coefficients and ε_i is the residual constant.

Equation 25 shows that y can be predicted using a range of x_n descriptors and multiplying them by coefficients, β_n. If there are more samples than variables, a solution can be found by minimising the residual, ε_i (Kowalski, et al., 1986).

However, this technique is prone to overfitting and there may be problems caused by collinearity. Collinearity occurs when there is a relationship between two parameters being used to build the model. This can be identified by a variance inflation factor (VIF) greater than 10. Overfitting describes a well fitted model with no predictive ability and can be avoided by assessing the significance of each parameter to the model individually (Roy, et al., 2008).

1.8.3.1.2. Partial Least Squares (PLS)

An alternative to MLR is Partial Least Squares (PLS) which combines MLR and Principal Component Analysis (PCA), giving it the advantage that it is not affected by collinearity. It is based on the same principles as MLR, using multiple X variables to predict Y. The variables X and Y can be modelled separately to give a relationship between them, however a more accurate model is produced when each model is given information about the other.

$$X = \sum_a t_a p_{ak} + e_k$$

Equation 26 - X is a variable, t_a is the X-score, p_{ak} is the loading and the residual by e_k.

The latent variables (Equation 26) are estimated using orthogonal X scores (t_a) where a = 1, 2, ... , A and A is small. The X-scores are linear combinations of the original x_k variable. They are weighted using coefficients $w_k{}^*{}_a$ where a = 1, 2, ... , A. A scores plot can be used to visually identify correlation between observations.

$$t_a = \sum_k w^*_{ka} X_t$$

Equation 27- t_a is the X-score, w^*_{ka} is the weighting coefficient which are related to X variable, X_t.

The loadings p_{ak} are aimed at providing good summaries of X that give the smallest value of e_k. Similarly the Y variables are calculated using weighted Y-scores, loading values to give a small residual. A loading plot can be used to visually identify correlations between variables. When compared to a scores plot, correlations between variables and observations can be found.

$$Y_m = \sum_a u_a c_{am} + g_m$$

Equation 28 - Y_m is the Y variable, u_a, is the Y-score, c_m is the weighting coefficients relating of the Y variables and g_m is the smallest residual.

These relationships take X and Y as separate entities and are known as the “outer relations”. If there is a relationship between X and Y, the X-scores will be good predictors for Y. This is known as the “inner relation”.

$$Y_m = \sum_a c_{ma} t_a + f_m$$

Equation 29 - the Y variables are denoted by Y_m, the weighting coefficients by c_m, the X-scores by t_a and new residual f_m.

Inner relations can be used to visually examine the relationship between X and Y using plots of t_1/t_2 and t_2/u_2.

$$Y_{im} = \sum_a c_{ma} \sum_k w^*_{ka} x_{ik} + f_{im} = \sum_k b_{mk} x_{ik} + f_{im}$$

Equation 30 - Y_{im} is the Y variable, c_{ma} is the weighting coefficients relating of the Y variables, w^*_{ka} is the weighting coefficients relating of the X variables, x_{ik} is the X variable, f_{im} is the residual, and b_{mk} is the inner relation

The inner relation (Equation 30) leads to the PLS equation (Equation 31).

$$b_{mk} = \sum c_{ma} w_{ka}^{*}$$

Equation 31 - b_{mk} is the inner relation, c_{ma} is the weighting coefficients relating of the Y variables, w_{ka}^{*} is the weighting coefficients relating of the X variables.

From Equation 30, it can be seen that PLS regression coefficents are equal to the weighted sum of the scores.

1.8.3.2. Machine Learning Techniques

Machine learning techniques utilise algorithms that learn from given data to make predictions. The AutoQSAR platform uses Random Forest (RF) and Support Vector Machines (SVM). It is an AstraZenca platform that originally included PLS and RF (Wood, et al., 2011) then Bayesian Neural Networks (Davis, et al., 2013) and now SVM. However, other examples of AutoQSAR models can be found in the literature (Rodgers, et al., 2011). Automatic QSAR model generation greatly decreases the time taken to produce a QSAR model. An AutoQSAR can employ many different mathematical methods. The AutoQSAR platform for this research employed PLS, RF and SVM. The software used by the AutoQSAR platform is R's RF Random Forest algorithm 4.6-6 and "A Library for Support Vector Machines (LIBSVM)" for the Support Vector Machines modelling (Chang, et al., 2011).

1.8.3.2.1. Random Forest (RF)

The Random Forest method is based around decision trees which are models that show the consequences of a decision. A starting node is split into two more nodes by a decision which are each in term spilt by a decision. Grouping a large number of these decision trees together to form an ensemble provides many different predictions. If the prediction is a categorical prediction, the predicted category from the majority of the trees is taken as the overall predicted category. If the predictions are numerical values, an average of the predictions is used as the predicted value. By using a large number of trees, the likelihood of reaching the correct result is increased. The growth of each tree is overseen by a randomly generated vector. There are several ways to improve the accuracy of a decision tree model. By randomly selecting the training set multiple times and averaging the results, accuracy is increased. The three most popular methods for identifying the training set are bagging (Breiman, 1996), random split selection (Dietterich, 1998) and boosting (Schapire, et al., 1998). Introducing

randomness into the selection, which is designed to even out anomalies, increases the likelihood of the correct solution being obtained. Breiman introduced another level of randomness by randomly selecting a subset of the variables that is significantly smaller than the total number of variables to be used as the training set. This is known as Random Forests and has the added benefit of reducing the computational expense of the calculation (Breiman, 2001).

1.8.3.2.2. Support Vector Machines (SVM)

Support Vector Machines (SVM) is a machine learning method that was pioneered by Vapnik (Vapnik, et al., 1963). The first step in the prediction is to split the training data set into two categories in such a way as to maximise the distance between the two categories. A median line is drawn between these two categories which are used to determine the category of the test set data by which side of the line the test data falls. The "Support Vectors" are drawn at the limit of each category.

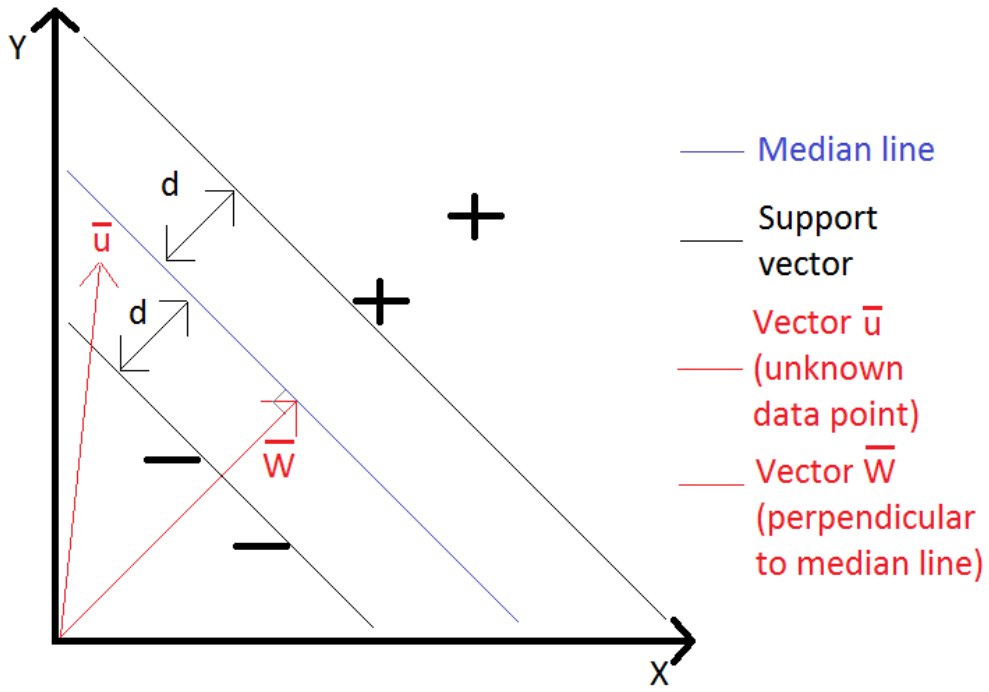


Figure 10 – Positive (+) and negative (-) data divided into two regions (categories) with their edges defined by two support vectors (black lines). The median line (blue) is half way in between the two support vectors. Two vectors of unknown length are indicated by red lines with vector $\bar{W}$ perpendicular to the median line between the two support vectors and vector $\bar{u}$ leads to an unknown point being predicted.

Figure 10 shows two sets of data, positive (+) and negative (-), with two support vectors (black lines). These support vectors are selected to give the widest gap (d) possible between the edges of the two data sets. Vector $\bar{W}$ is perpendicular to the median line between the two support vectors and is of unknown length. The vector $\bar{u}$ leads to an unknown point being predicted. Whether the vector $\bar{u}$ will fall on the positive or negative side of the median line is dependent on its length. If the length of vector $\bar{u}$ is larger than the distance between the axis and the median line (C), it will fall into the positive category. If the distance is smaller, the point being predicted is in the negative category. The dot product of the vectors $\bar{u}$ and $\bar{W}$ provides the length to be compared to length, C (Equation 32).

$$\bar{W} . \bar{u} \geq C$$

Equation 32 - $\bar{W}$ is the vector which is perpendicular to the median line between the two support vectors of unknown length, $\bar{u}$ is the unknown vector and C is a constant.

This equation can be simplified by taking C = -b to produce Equation 33. If Equation 33 is true (i.e. the length of vector $\bar{u}$ is longer than the median line), the end of vector $\bar{u}$ falls with the category containing positive data and if false (i.e. the length of vector $\bar{u}$ is smaller than the median line) falls within the category containing negative data.

$$\bar{W} \cdot \bar{u} + b \geq 0$$

Equation 33 – Decision Rule where $\bar{W}$ is the vector which is perpendicular to the median line between the two support vectors of unknown length, $\bar{u}$ is the unknown vector and b is the constant.

The length at which vector $\bar{u}$, meets the median line is set as zero. To account for error in the calculation of the length of vector $\bar{u}$, the assumption that a length between 0 and 1 is too close to the median line to be certain that the data really is positive and length between 0 and -1 is too close to be sure that the data is really negative. This leads to two equations which define the regions in which positive (Equation 34) and negative data (Equation 35) will be found.

$$\bar{W} \cdot \bar{x}_+ + b \geq 1$$

Equation 34 - $\bar{x}_+$ indicates a positive sample, $\bar{W}$ is the vector which is perpendicular to the median line between the two support vectors of unknown length and b is the constant.

$$\bar{W} \cdot \bar{x}_{-} + b \leq -1$$

Equation 35 - $\bar{x}_{-}$ indicates a negative sample, $\bar{W}$ is the vector which is perpendicular to the median line between the two support vectors of unknown length and b is the constant.

Equation 34 and Equation 35 can be combined using y_i such that y_i = +1 for positive data and y_i = -1 for negative data. Both equation produce Equation 36.

$$y_i(\bar{W} \cdot \bar{x}_i + b) \geq 1$$

$$y_i(\bar{W} \cdot \bar{x}_i + b) - 1 = 0$$

Equation 36 - y_i indicates the category of the sample, $\bar{x}_i$ is the sample vector, $\bar{W}$ is the vector which is perpendicular to the median line between the two support vectors of unknown length and b is a constant.

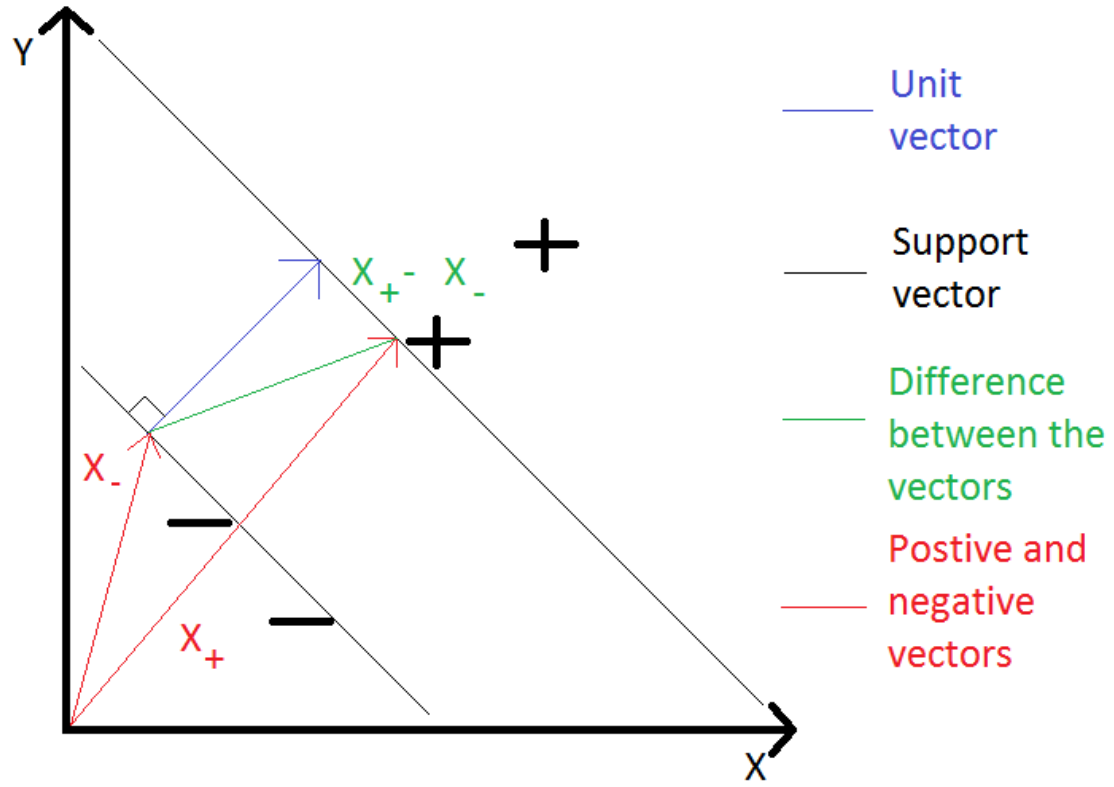


Figure 11 – Positive (+) and negative (-) data divided into two regions (categories) with their edges defined by two support vectors (black lines). The difference between the positive and negative vectors (red lines) is donated by the green line allowing trigonometry to be used to calculate the distance between the two support vectors known as the unit vector (blue line).

The distance between the two support vectors can be calculated using the difference between the edges of the two categories (positive vector, $\bar{x}_+$ and negative vector, $\bar{x}_-$) (Figure 11 and Equation 37).

$$distance\ between\ support\ vectors = (\bar{x}_+ - \bar{x}_-) \cdot \frac{\bar{W}}{||W||}$$

Equation 37 - $\bar{x}_+$ is the positive vector, $\bar{x}_-$ is the negative vector and W is the vector which is perpendicular to the median line between the two support vectors of unknown length.

The first part of Equation 37, $(\bar{x}_+ - \bar{x}_-)$ is also known as the dot product which is scalar. The second part of Equation 37 is to turn the original $\bar{W}$ from Figure 10 into the unit vector by dividing it by its magnitude as Vector $\bar{W}$ is perpendicular to the median line. Equation 36 is substituted into the first part of Equation 37 for positive data to give Equation 38.

$$\bar{x}_+\bar{W} + b - 1 = 0$$

$$\bar{x}_+\bar{W} = 1 - b$$

Equation 38 - $\bar{x}_+$ is the positive vector, W is the vector which is perpendicular to the median line between the two support vectors of unknown length and b is a constant.

Alternatively, for negative data when Equation 36 is substituted into the first part of Equation 37, Equation 39 is formed.

$$-\bar{x}_-\bar{W} - b - 1 = 0$$

$$\bar{x}_-\bar{W} = -1 - b$$

Equation 39 - $\bar{x}_-$ is the negative vector, W is the vector which is perpendicular to the median line between the two support vectors of unknown length and b is a constant.

Therefore substituting Equation 38 and Equation 39 into Equation 37 leads to Equation 40 which is the distance between the support vectors.

$$distance\ between\ support\ vectors = \frac{\bar{x}_-\bar{W} - \bar{x}_-\bar{W}}{\left|\left|W\right|\right|} = \frac{(1-b) - (-1-b)}{\left|\left|W\right|\right|} = \frac{2}{\left|\left|W\right|\right|}$$

Equation 40 - $\bar{x}_-$ is the negative vector, b is a constant and W is the vector which is perpendicular to the median line between the two support vectors of unknown length

The two categories that the data is initially spilt into are specifically chosen to maximise the distance between them (i.e. the support vector as far away from each other as possible) and therefore the width (2d) to be as large as possible. Therefore, Equation

40 needs to be maximised. The constant (2) is small enough to be disregarded and Equation 40 is converted to its reciprocal to be minimised (Equation 41).

$$width = \left|\left|W\right|\right| = \frac{1}{2}(||W||)^2$$

Equation 41 - W is the vector which is perpendicular to the median line between the two support vectors of unknown length.

Equation 41 is then used in combination with a Lagrunge multiplier. This allows the calculation of the extreme minima with constraints and gives rise to Equation 42 that can be used to calculate the maximum and minimum without taking constraints into account.

$$width\ with\ Lagrunge\ multipler\ (L) = \frac{1}{2}(||W||)^2 - \sum \alpha_i[y_i\ (\bar{W}\ .\bar{x}_i\ +\ b) - 1]$$

Equation 42 - where $\bar{x}$ is the sample vector, W is the vector which is perpendicular to the median line between the two support vectors of unknown length, a_i is the Lagrange multiplier and y_i indicates the category of the sample.

Equation 42 calculates the width by taking the function to be minimised (Equation 41) and subtracting the summation over all of the constraints where alpha is the Lagrunge multiplier and the constraint is Equation 36. As the aim is to minimise the width, the derivatives need to be calculated and set equal to 0 (Equation 43).

$$\frac{dL}{d\bar{W}} = \frac{2}{2}\bar{W}^1 - \sum \alpha_i y_i\, \bar{x}_i = 0$$

$$\bar{W} = \sum \alpha_i y_i\, \bar{x}_i$$

Equation 43 - where $\bar{x}$ is the sample vector, W is the vector which is perpendicular to the median line between the two support vectors of unknown length, a_i is the Lagrange multiplier and y_i indicates the category of the sample.

This is done using partial differentiation with respect to $\bar{W}$ gives linear sum of the samples (Equation 43) and with respect to b (Equation 44).

$$\frac{dL}{db} = -\sum \alpha_i y_i = 0 = \sum \alpha_i y_i$$

Equation 44 - L is the length, W is the vector which is perpendicular to the median line between the two support vectors of unknown length, a_i is the Lagrange multiplier and y_i indicates the category of the sample.

Equation 43 can be substituted into Equation 42 to give Equation 46 which is made up of two dot products followed by two constants.

$$L = \left[\frac{1}{2}\left(\sum \alpha_i y_i \bar{x}_i\right) \cdot \left(\sum \alpha_j y_j \bar{x}_j\right)\right] - \left[\sum \alpha_i y_i \bar{x}_i \cdot \left(\sum \alpha_j y_j \bar{x}_j\right)\right] - \sum \alpha_i y_i b + \sum \alpha_i$$

Equation 45 - L is the length, $\bar{x}$ is the sample vector, a is the Lagrange multiplier, b is the constant and y indicates the category of the sample.

$$L = -\frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i \cdot x_j - 0 + \sum \alpha_i$$

Equation 46 - L is the length, $\bar{x}$ is the sample vector, a is the Lagrange multiplier and y indicates the category of the sample.

$$\sum \alpha_i y_i x_i \cdot \bar{u} + b > 0 \text{ if positive}$$

Equation 47- u is unknown vector, a_i is the Lagrange multiplier, y_i indicates the category of the sample, x_i is the sample vector and b is the constant.

In the example used in Figure 10 and Figure 11, the two categories of data can be easily linearly separated. However, if the data cannot be linearly separated then a transformation $\phi(\bar{x})$ can be applied to allow the data to be linearly separated. Equation 48 need to maximise as Equation 46 is only dependant on a dot product of two vectors.

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$$

Equation 48 - K is the Kernel function x is the sample vector and ϕ is the transformation.

A Kernel function provides the dot product of those two vectors in another space (Equation 48). Therefore, it is not necessary to know the transformation used to make the data linearly separable. The Kernel function that is used in the AutoQSAR software is the radial basis function (RBF) as shown in Equation 49.

$$K\left(x_i, x_j\right) = e^{\left(-\gamma||x_i - x_j||^2\right)}$$

Equation 49 – Radial Basis Function where K is the Kernel function, x is the sample vector and gamma is greater than zero.

Once the Kernel function that maps the training set data has been established, it can be used to predict an unknown y from a known x value using Equation 50.

$$y(x) = \alpha_i K\left(x_i, x_j\right) + b$$

Equation 50 - y is the value to be predicted, a_i is the Lagrange multiplier, K is the Kernel function, x is the sample vector and b is a constant.

SVM can be run using many different types of software (such as LIBSVM) and can be used in conjunction with many validation techniques, such as cross validation. Czermiński et al. demonstrated that SVM models outperform those generated using NN for highly non-linear data when generating QSAR models (Czermiński, et al., 2001).

1.8.4. 3D QSAR modelling

Comparative Molecular Field Analysis (CoMFA) allows the chirality of a compound to be taken into account by constructing 3D QSAR models (Cramer III, et al., 1988). It works by calculating a low energy conformation of the molecule using atomic partial charges. A box is then placed around the molecule and used to create a grid. A field value is assigned to each grid point, which are used as the molecular descriptors. Partial Least Squares (PLS) is used to generate the QSAR model by correlating the fields and biological activity and produce a regression equation (Kubinyi, 1998). This has applications in drug design as one chiral form of a molecule may be therapeutically active whilst the other can be toxic. However, it has also been used to develop a chiral stationary phase for use with HPLC (Schefzick, et al., 2000).

Another approach to developing QSAR models is Fragment-Based QSAR model (FB-QSAR). This is particularly useful for in silico drug design where one or two functional groups are very important. QSAR models need to be able to predict how small structural changes can affect biological activity. FB-QSAR model has the added functionality of being able to weight fragments by combining a free energy coefficient with Free-Wilson analysis. An iterative double least squares method can then be used

to solve the 3D linear equations for two sets of coefficients (Du, et al.). The method of Fragment-Based QSAR model has further been developed to the Hologram QSAR model. This method converts the fragments into a molecular “fingerprint” known as a hologram. PLS is then used to identify the fragments that have the most significant effect of the descriptors. It can also be used to give an indication of the numbers of each fragment within a data set (Salum, et al., 2009). The process can be reversed (Inverse QSAR) to generate a molecule descriptor which allows the prediction of a compound with the ideal properties (Wong, et al., 2009).

1.8.4.1. Predicting Partition Coefficients using QSAR models

There have been many examples of QSAR models being used to predict partition coefficients. Traditionally, these models have been developed for much simpler two solvent systems such as octanol/water and cyclohexane/water (Abraham, et al., 1996). Octanol/water partitioning has been of particular interest to the pharmaceutical industry as a model of the blood brain barrier (BBB). The first prediction of octanol/water partition coefficients (logP) was a fragment based method using a large experimental data set. Rekker and Nyes used a large number of fragments with few correction factors to predict logP values (Nys, et al., 1974). Leo and Hansch further developed this fragment based approach using a small number of fragments combined with a large number of correction factors. This approach led to an automated prediction tool called “Calculated octanol/water partition coefficient (ClogP)” which made use of a database of more than 30,000 values (Leo, et al., 1975). Since the development of ClogP, there have been many automated logP prediction tools developed including other fragment based approaches including the Available Chemical Directory octanol/water partition coefficient (ACDlogP). This software uses an algorithm to estimate any missing fragments. However, it is based on a data set which contains 3601 compounds which is considerably smaller than the data set used in ClogP (Petrauskas, 2000). An alternative to the fragments based approached are atom based approaches. Ghose and Chippen produced an atom based fragment approach which could be used as GClogP (Ghose, et al., 1986) or applied to artificial neural networks (ANN) to produce the NNlogP values. There are many other approaches, with Livingston naming 19 commercially available software packages (Livingstone, 2003).

QSAR models have also been used to predict partitioning between water and alkanes ($logP_{alk}$), although a lack of a large, diverse and consistent database had prevented the development of a predictive tool like ClogP. Abraham applied a general solvation equation of a linear solvation energy relationship (LSER) to water/hexadecane (Abraham, et al., 1990) and compared these partitions to water/alkane, water/octanol and water/cyclohexane (Abraham, et al., 1994). A "critical quartet" of partitioning systems was established by Leahy et al. (Leahy, et al., 1992) to cover partitioning between water and an inert solvent, an amphiprotic solvent, a proton donor and a proton acceptor. The systems chosen for the water/inert solvent was water/alkane which included hexane, heptane and octane. The partitioning was modelled on LSERs. Taft and Kamlet had shown that the significant parameters in an LSER could be used to identify the chemical factors that were governing the partition in that particular system (Kamlet, et al.). Atom type models have been built using polar surface area (PSA) (Platts, et al., 2004), volume (Zerara, et al., 2009), electrostatics (Lamarche, et al., 2004), interaction fields (Caron, et al., 2005) and polarity (Kenny, et al., 2013).

1.8.4.1.1. Distribution coefficient (logD)

As the partition coefficients of ionised and neutral compounds differ, logP is not always an appropriate measure of lipophilicity for ionisable compounds. This is because only neutral species will have the tendency to partition into the organic layer. An ionisable compound is in constant dynamic equilibrium between its neutral and ionised form (Equation 51).

$$HA \rightleftharpoons A^- + H^+$$

Equation 51 – The acid dissociation equation with the acid (HA) dissociates to form a proton (H^+) and a negative species (A-).

The acid dissociation constant describes the tendency of a compound to ionise, with a higher K_a value indicating the compound is likely to ionise making it a stronger acid (Equation 52).

$$K_a = \frac{[A^-][H^+]}{[HA]}$$

Equation 52 – The acid dissociation constant (K_a) can be calculated using the concentrations of the acid (HA), the protons (H^+) and the negative species (A-).

These K_a values are very small so are often converted to the pK_a of a compound, by taking the negative logarithm of the K_a value. The lower the pK_a value of a compound, the stronger acid it is (Equation 53).

$$pKa = -\log_{10} K_a$$

Equation 53 – The pKa of a compound is equal to the negative log of the acid dissociation constant (K_a) of a compound.

When the concentration of the neutral form of the molecules is equal to the concentration of the negatively charges species the pH is equal to the pKa.

The distribution coefficient (logD) is similar to logP but instead of using an octanol/water system, a buffer at a specific pH replaces the water. The logD of a monoprotic acid can be calculated using Equation 54 (Lombardo, et al., 2001).

$$\log D = \log P + \log\left(\frac{1}{1 + 10^{pk_a - pH}}\right)$$

Equation 54 - logD is the distribution coefficient, logP is the partition coefficient, pk_a is the acid dissociation constant and pH is the negative log of the concentration of hydrogen ions

As can be seen in Figure 12, at the point where the pKa of a compound is equal to the pH, the gradient of the line changes. When the pH of the system is greater than the pKa of the compound the logD value of the compound is pH dependent. Below this point the logD value is independent of pH change.

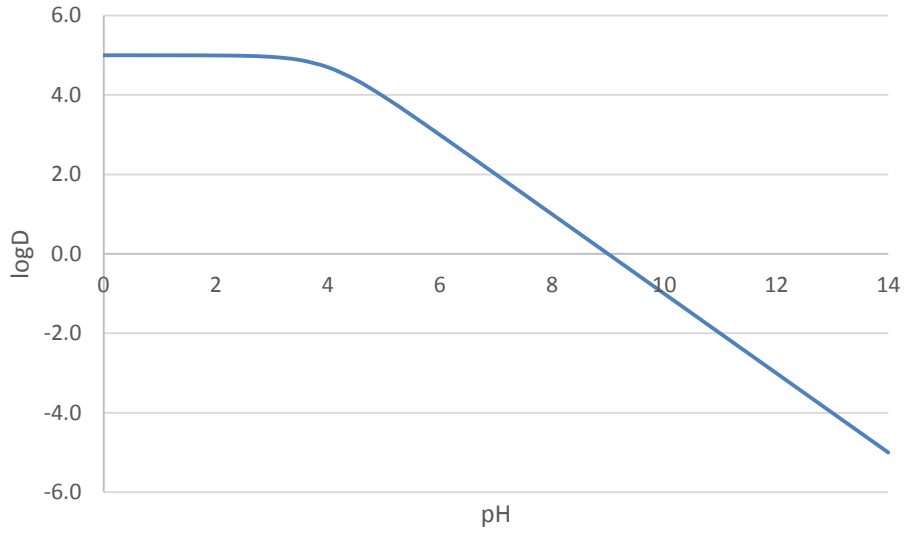


Figure 12 – A graphical representation of Equation 54 for a compound with logP of 5 and a pKa of 4.

QSAR models have been developed for the prediction of logD. There is commercially available software to predict logD. ACD labs have the ability to predict logD along with SLIPPER (Solubility, LIPophilicity, PERmeability) (Raevsky, et al., 2002) (TimTec LLC, Newark, DE, USA) and PrologD (Csizmadia, et al., 1997) (CompuDrug, Florida, FL, USA).

1.8.5. Assessing the predictive ability of a QSAR model

The success of a QSAR model is entirely dependent on the training set data used to build it. This data must be as accurate as possible since the experimental error will be compounded by the computational error in the QSAR model. There is a risk to combining experimental data from multiple sources as the accuracy of the results cannot be guaranteed. The development of high-through-put techniques has allowed a huge growth in QSAR usage (Selassie, et al., 2010). The training set needs to span parameter space as widely as possible, giving the best possible likelihood that a test molecule will fall with the measured parameters space. This also increases the tolerance of the model to errors within the experimental data. By distributing the training set evenly, the amount of experimental work is reduced whilst ensuring cross validation is still applicable and anomalous experimental training set data do not have a detrimental effect on the QSAR model (Kubinyi, 1998).

Methods for assessing the predictive ability of a QSAR model can be broadly assigned into two categories: external validation and internal validation. These are very

important when it comes to proving that a QSAR model is reliable and accurate (Gramatica, 2007). External validation involves testing the model using a test set that was not used to train the model. Internal validation involves leaving out part of the training set to be used as a test set. Roy and Roy compared four methods of assessing the predictive ability of a QSAR model (Roy, et al., 2008). Splitting the data into test and training sets is generally considered the most vigorous method of checking the QSAR model as this is the closest model to the real life use of the QSAR model once is has been generated (Dearden, et al., 2009). Roy and Roy concluded that the most important factor in the success of a QSAR model was how well the initial data set covered parameter space (Roy, et al., 2008).

Once the models have made the predictions, these must be compared to the experimental values to determine whether the predictions are acceptable. The simplest way to calculate the difference (Δ) between them. The closer the Δ value is to zero the better. For this work, the Δ value between the experimental and predicted values must be less than 0.5 to be deemed acceptable.

1.8.5.1. Cross validation (CV)

Multivariate Analysis (MVA) techniques work on the assumption that not all of the available data is useful when modelling a relationship. Variables are removed whilst a correlation of patterns is observed for the data, removing the noise from the model. However, at some point removing variables reduces the predictive ability of the model and is therefore counterproductive. Cross validation (CV) is used to identify this point by repeatedly calculating the coefficient of determination (R^2) for the data as one value from the data set is left out of this calculation. R^2 is a measure of how well the regression model accounts for variation in the experimental data. The R^2 value ranges from 0 to 1 with 0 being a very poor fit with the data and 1 being a perfect fit with the data. The QSAR model can then be used to predict the value that was left out. This predicted value is then compared to the left out value. The accuracy of the model can then be assessed by how close the predicted value is to the experimental value. When repeated for every value in the data set, this method provides an overview of the predictive ability of the model. However, this validation method can be time consuming.

1.8.5.2. Divide the data into training and test sets.

This involves removing a proportion of the data from the building process of the QSAR model denoted the “test set”. The remaining data, known as the “training set”, is then used to build a QSAR model. The subsequent QSAR model is then used to predict the values of the test set and the experimental and predicted values compared. This method can be dependent on how the data set is divided into the training and test sets. It is possible to add bias to the model by not covering the entirety of parameter space with the training or test sets.

1.8.5.3. Application of the model to external data.

This method generates a QSAR model from all of the available data is used. The QSAR model then makes predictions which are subsequently tested experimentally. This is the best practise method of testing a QSAR model as it avoids adding bias into the way that the data set is divided up into the training and test sets.

1.8.5.4. Data randomising

Data randomising is used to establish whether the model has genuine predictive ability or is the result of chance correlation between the descriptors and the model. This can be quantified using or the predictive squared correlation coefficient (Q^2). To calculate the Q^2 value of a prediction model, the data is split into seven groups. Six of the groups are then brought together to form a training set used build a model, which is tested by predicting the values in the remaining seventh group. This is repeated with each group being used as a test set. The differences between the predicted and experimentally determined values are squared and summed for the seven cross validations to calculate the predicted residual sum of squares (PRESS) (Equation 55).

$$PRESS = \sum_{i=1}^{n} (y_i - y_{i-1})^2$$

Equation 55 – The predicted residual sum of squares (PRESS)

Q^2 is then calculated using Equation 56.

$$Q^2 = 1 - \frac{PRESS}{\sum y^2}$$

Equation 56 – The predictive squared correlation coefficient (Q^2).

The higher the Q^2 value, the greater the predictive ability of the model. If a Q^2 value is above 0.65, it is considered a good QSAR model (Umetrics, 2015).

1.8.6. Statistics

The models will be compared using R^2, Q^2 and Root Mean Square Error (RMSE) values. An R^2 value above 0.78 is considered a good QSAR model. The RMSE value is calculated based on the difference between the predicted and actual values (Equation 57). The lower the RMSE the better the fit of the model with an RMSE of less than 0.5 being set as the acceptance criteria.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(x_i - y_i)^2}{N}}$$

Equation 57 - Root Mean Square Error where x_i is the experimental value with y_i as the corresponding predicted value and N is the total number of observations.

A Q^2 value above 0.65 is considered a good QSAR model.

The significance of a descriptor can be assessed using *p*-values. This is "an estimated probability of rejecting the null hypothesis (H_0) of a study question when that hypothesis is true" with a lower the value suggesting a more significant descriptor.

1.9. Conclusion

CCC/CPC are an alternative to many solid-liquid techniques. However, the available methods for solvent system selection are less than ideal. The methods that require experimental measurements are impractical for use in the modern world of bio-pharmaceutical industry with expensive compounds or compounds of which there is only a very small quantity. The current computational models are complicated, involving long run times with a high level of knowledge required to run them, or involving a large capital investment in software before they can be run. This is prohibitive for a novice and is slowing CCC/CPC's transition to becoming a mainstream analytical/research techniques. QSAR models have been successfully applied to partitioning in solvent systems previously, making them a good candidate for the generation of a novel computational model.

2. Materials and methods

2.1. Chapter 3 - Investigating the factors that affect partition coefficient (K_d) values

2.1.1. Materials

The compounds to be investigated within this chapter were phenanthrene and 2-ethylanthraquinone which were purchased from Sigma Aldrich (Gillingham, UK). They were selected due to their high solubility in HEMWat and the fact that they have measurable K_d values in all six of the HEMWat systems. The HEMWat systems investigated were 8, 14, 17, 20, 22 and 26 prepared using the HPLC grade solvents heptane, ethyl acetate and methanol purchased from Fisher Chemicals (Loughborough, UK). The water used was deionised in house using a Purite Select Fusion purification system (Thame, UK). The additional solvents used were HPLC grade acetonitrile, dimethyl sulfoxide (DMSO) and ethanol along with the trifluroracetic acid (99%) and 1% ammonium hydrate (35%) which were purchased from Fisher Chemicals (Loughborough, UK). HPLC analysis was conducted on a HP1100 Agilent system (Stockport, UK) with detection at 254, 260, 275, 295, and 310 nm with a Symmetry C18 column (75 × 4.6 mm I.D., 3.5 µm), (Waters, USA). An Eppendorf Concentrator 5301 (Hamburg, Germany) was used as a centrifuge at 1400 rpm (240g) at room temperature (20°C). The balance used was a Sartorius Mechatronics analytical balance 1601A MP8-1 (Epsom, UK) unit with a range from 0.1 mg to 110 g. The CCC centrifuge used was the “Dynamic Extractions Mini” with a rotor radius of 5 cm and PTFE tubing with an internal diameter 0.8 mm (Slough, UK). A HP1100 Agilent system (Stockport, UK) with detection at 254, 260, 275, 295, and 310 nm was used as the pump and for detection.

2.1.2. The method to assess the effect of the addition of co-solvent

Initially, 10 mg of phenanthrene and 2-ethylanthraquinone were weighed into separate vials and dissolved in 1.5 ml of the initial solvent (either DMSO or the upper phase of the HEMWat system being investigated). The solution was sonicated to ensure that the entire compound had been dissolved. The four HEMWat systems were made up according to Table 1 by volume. These were; HEMWat 14, 17, 20, and 22. Once these had been left to settle into two phases and equilibrate for 15 minutes, 30 µl of the solution containing the compound was added to a HPLC vial containing 600 µl of each

phase of the solvent system. The vials were then vortex mixed and centrifuged. The upper phase was sampled and the remaining upper phase was removed using a pipette so that only the lower phase was left in the vial. A 250 µl sample was taken from the lower phase, free from contamination with the upper phase. This was done in triplicate leading to three samples that were run on a 10 minute gradient method on the HPLC using Symmetry C18 column (4.6x75mm, 3.5um), at 1ml/min and 40°C. The mobile phase consisted of 0.1% aqueous trifluoroacetic acid (solvent A) and acetonitrile (solvent B). The gradient elution program was as follows: 0-6 min, 10% B; 2-8 min, 80% B with a total run time of 10 minutes.

The peak areas of the components within each phase were obtained through integration of the HPLC chromatogram. The K_d value was then determined by dividing the peak area of the upper phase by the peak area of the lower phase.

2.1.3. The method to assess the effect of temperature

Six HEMWat systems were prepared according to Table 1 by mass rather than by volume. This was to eliminate any effect that temperature could have on the volume of the solvents and therefore affect the percentage solvent compositions of the HEMWat systems. The six HEMWat systems were; HEMWat 8, 14, 17, 20, 22 and 26. The vials containing the solvent systems were added to a water bath and warmed to the required temperature. Initially, 10 mg of compound was dissolved in 1.5 ml of the upper phase of the relevant HEMWat system. The partition coefficient was measured using the same method used in section 2.1.2.

2.1.4. The method to assess the effect of equilibration time

The solvent systems were made up by adding the correct ratio of each of the four solvents to a vial according to Table 1 by mass. The solvents were preheated to 20°C in a water bath. The system was then shaken and left to equilibrate. Repeated K_d measurements were made at five different time points. When there was no longer a change in the K_d values observed, it was concluded that the system had reached equilibrium. The five different equilibration times were:

1. 0 minutes i.e. the solvent system was made, shaken and as soon as had separated into two phases, it was sampled for K_d measurement.

2. 30 minutes i.e. the solvent system was made, shaken, allowed to form two phases. It was then left to sit in a 20°C water bath for 30 minutes before it was sampled.

3. 1 hour i.e. the solvent system was made, shaken, allowed to form two phases. It was then left to sit in a 20°C water bath for 1 hour before it was sampled.

4. 2 hours i.e. the solvent system was made, shaken, allowed to form two phases. It was then left to sit in a 20°C water bath for 2 hours before it was sampled.

5. Overnight i.e. the solvent system was made, shaken, allowed to form two phases. It was then left to sit in a 20°C water bath overnight before it was sampled.

The K_d values were obtained using the method from section 2.1.2.

2.1.5. The method to assess the effect of solute concentration

Each of the four solvents (heptane, ethyl acetate, methanol and water) was placed into a water bath at 20°C before they were used to make up the solvent systems. Six solvent systems were made by mass according to Table 1 and replaced into the water bath and left overnight to equilibrate. The compound to be studied was weighed into HPLC vials. The solvent systems were removed from the water bath and sampled. One phase (1.5 ml) was saturated with the compounds being investigated. The phase was deemed saturated if after 1 hour, there was undissolved compound present at 20°C. To make sure that there were no particulates in the supernatant, the HPLC vials were centrifuged for 1 minute at 1400 rpm at room temperature (20°C). Once this had been completed, aliquots of 400 µL of the supernatant were pipetted into HPLC vials and 1400 µl of the alternative phase was added into the vial by pipette. The vials were then vortex mixed for 30 seconds and centrifuged for 1 minute at 1400 rpm at room temperature (20°C). An aliquot of 80 µl of the 1400 µl volume phase was pipetted into 1 ml of ethanol and 320 µl of the 400 µl phase was pipetted in 1 ml of ethanol. Before the lower phase was sampled the remaining upper phase was removed and discarded using a pipette. This was done in triplicate for each of the six HEMWat systems. The samples were run on a 10 minute gradient method on the HPLC. The HPLC method details are shown in Table 28. Once this value had been calculated the sample was

diluted by pipetting 750 µl of the saturated solution to 750 µl of fresh HEMWat. This was repeated four times to give four dilutions.

2.1.6. The method to assess the effect of pH

All the CCC experiments were carried out in reverse phase (the upper, less polar phase was used as the stationary phase) with a flow rate of 1 ml/min. The solvent systems were prepared in two different ways. The first method involved preparing the HEMWat systems by mass using the ratios in Table 1 and leaving them to equilibrate overnight. The phases were separated and each was sonicated for 15 minutes before being pumped into the system. The alternative was to make the systems using mixing on demand with the quaternary pump on the HP1100 Agilent system (Stockport, UK) used to make up each phase of the system. The percentage compositions of the upper and lower phases can be found in Table 2 and Table 3.

Table 2 – Percentage composition of the upper phase of the six HEMWat systems when prepared by volume.

HEMWat number	Percentage Heptane	Percentage Ethyl Acetate	Percentage Methanol	Percentage Water
8	10	87	1	2
14	37	59	3	1
17	63	34	2	1
20	85	14	1	0
22	91	8	1	0
26	97	2	1	0

Table 3 – Percentage composition of the lower phase of the six HEMWat systems when prepared by volume.

HEMWat number	Percentage Heptane	Percentage Ethyl Acetate	Percentage Methanol	Percentage Water
8	0	8	9	83
14	0	13	28	59
17	0	19	41	40
20	1	18	54	27
22	1	15	63	21
26	4	7	80	9

The sample containing the mixture of compounds was dissolved one of the phases of HEMWat system depending on the molecules' nature. The concentration was solubility dependent with a maximum of 20 mg/ml of each compound. The experiment was conducted at 20°C with a 20 µl injection volume and a rotation speed of 2100 rpm.

2.2. Chapter 4 - Generating the QSAR models

2.2.1. Method used to generate for QSAR models using Partial Least Squares (PLS)

SIMCA-P version 13 (Umetrics, Umea, Sweden) was used to perform the PLS regression. The initial QSAR was generated using the software tool "Autofit" which carried out a PLS regression using all 196 descriptors (listed in section 9.3). The significance of the descriptors for the original model was assessed using the Variance Inflation Plot (VIP). A VIP value is calculated for each descriptor (x_k) using the sum of the squares of the PLS loadings weights (wa_k) weighted by the amount of the sum of squares explained in each model component. A descriptor with a VIP value greater than 1 indicates that it is significant. Conversely, a descriptor with a VIP value of less than 0.5 is insignificant. For any descriptor with VIP values that are between these two values it is less clear whether they are significant or not. In this case, there are more descriptors than data points, therefore the threshold for significance was set at 1 being the lowest value considered to indicate importance. Accordingly, any descriptors with a VIP value of less than 1 were removed from the original model and the remaining

descriptors were used to generate a second PLS model, again using the “Autofit” tool. Once this model had been built, the same process of removing descriptors with a VIP value of less than 1 was repeated and a third PLS model was built. For four of the six HEMWat systems, this third model proved to be poorer than the second. This is most likely due to the low number of descriptors used to build the models as any descriptors with a VIP value of less than one had been removed twice. Therefore, a fourth model was not built. The Root Mean Square Error (RMSE), the coefficient of determination (R^2) and the predictive squared correlation coefficient (Q^2) values were calculated and used to assess the models’ predictive ability. Once these three QSAR models from PLS had been generated, the process was repeated using the “Top 14” AstraZeneca descriptors only (listed in section 9.2). This procedure was carried out for each of the six HEMWat systems. This QSAR model was then used to predict the $\log K_d$ values of the four test compounds: biphenyl, benzoquinone, tolbutamide and quinine.

2.2.2. Method used to generate QSAR models using Multiple Linear Regression (MLR)

The Multiple Linear Regression (MLR) was carried out using JMP 10.0 (SAS Institute, Cary, NC, USA). The QSAR models were generated as follows:

a) All 196 descriptors and stepwise linear regression (listed in section 9.3).

b) “Top 14” AstraZeneca descriptors and stepwise linear regression (listed in section 9.2).

c) The two most important descriptors from the partitioning tool in JMP from all 196 the descriptors.

d) The two most important descriptors from the partitioning tool in JMP from the “Top 14” AstraZeneca descriptors.

e) Five Abraham parameters and stepwise linear regression (Taft, et al., 1985).

The QSAR models were generated for the training set using each of the five different combinations of descriptors. Each time the results where manually checked to guard against overfitting, using the *p*-values to assess their significance. A *p*-value is

considered statistically significant as if $P < 0.05$ and considered highly significant if $P < 0.001$. The fewer descriptors used to produce an MLR the better as this reduces the likelihood of overfitting. If more than three descriptors are identified by the stepwise regression then the descriptor with the largest *p*-value was removed and the regression repeated. Once only three descriptors remained, if they all had *p*-values below 0.05 they were considered significant. If not, another was removed until all the descriptors had a *p*-value of less than 0.05. This process was repeated for each of the five sets of descriptors. This gave five QSAR models produced using MLR. From these QSAR models, the R^2 and RMSE of the training set were compared. The QSAR model with the highest R^2 values and the lowest RMSE was selected as the best performing QSAR model. This QSAR model was then used to predict the $\log K_d$ values of the four test compounds.

2.2.3. Method used to generate QSAR models using Random Forest (RF)

The online AstraZeneca internal platform AutoQSAR was used to generate Random Forest (RF) models for each of the six HEMWat systems. It makes use of the RF software programme in R and automatically optimises the numbers of trees and subtrees. These settings are default on the AutoQSAR platform and cannot be altered.

2.2.4. Method used to generate QSAR models using Support Vector Machines (SVM)

The online AstraZeneca internal platform AutoQSAR was used to generate Support Vector Machine (SVM) models for each of the six HEMWat systems. This software makes use of the software LIBSVM with the Kernel function set as the radial basis function. These settings are default on the AutoQSAR platform and cannot be altered.

2.3. Chapter 5 - Testing the QSAR models

2.3.1. Materials

The compounds to be investigated within this chapter were purchased from Sigma Aldrich (Gillingham, UK) and were all of 99% purity or greater. The remaining materials can be found in section 2.1.1.

2.3.2. The method to carry out the initial testing of the models with CCC runs with HEMWat systems prepared by mass

All the CCC experiments were carried out in reverse phase (the upper, less polar phase was used as the stationary phase) with a flow rate of 1 ml/min. The solvent systems were made up by mass using the ratios in Table 1 and left to equilibrate overnight in a separating funnel. The phases were separated and each was sonicated for 15 minutes to remove dissolved gas before being pumped into the system. The total volume of the system was measured at 20 ml and included the pump and CCC system. The sample containing the mixture of compounds was dissolved in the upper phase of HEMWat system in which the mixture was to be separated, with a maximum of 20 mg/ml of each compound (solubility dependent). The experiment was conducted at 20°C with a 20 µl injection volume and a rotation speed of 2100 rpm.

2.3.3. The method to test the models with CCC runs with HEMWat systems prepared by mixing on demand

All the CCC experiments were carried out in reverse phase (the upper, less polar phase was used as the stationary phase) with a flow rate of 1 ml/min. The solvent systems were made up by mixing on demand. Each of the four solvents were attached to one of the four line on the quaternary pump and the phases were made up in the pump. The phases were made up using the percentage compositions in Table 2 and Table 3. The sample containing the mixture of compounds was dissolved in the upper phase of HEMWat system in which the mixture was to be separated, with a maximum of 20 mg/ml of each compound (solubility dependent). The experiment was conducted at 20°C with a 20 µl injection volume and a rotation speed of 2100 rpm.

3. Investigating the factors that affect partition coefficient (K_d) values

The accuracy of a QSAR model is wholly dependent upon the accuracy of the experimental data used to train it. However, the experimental accuracy of a measured partition coefficient (K_d) value can be affected by many factors. Controlling these factors is critical to ensuring the accuracy and reproducibility of the experimental K_d measurements. Before these physical factors were assessed, criteria for the materials used in the experiments were set out. With the aim of preventing any solute impurities from affecting the accuracy of the K_d measurement, it was decided to only use compounds with a purity above 99%. The same concern motivated the decision to use HPLC grade solvents along with HPLC grade water produced in house. As water can gain contaminants from the air or absorb gases, it was obtained just before the preparation of the solvent systems. It had also been noted in the literature that the pH of HEMWat systems is not stable over a long period of time (Sumner, 2011). It was suggested that this was due to the hydrolysis of ethyl acetate which had been known to change the pH of a HEMWat systems by up to 3 units, with the problem particularly affecting systems with low methanol content (Berthod, et al., 2005). For this reason, the HEMWat systems for each experiment were freshly prepared to ensure each measurement was conducted at a constant pH (Sumner, 2011).

Phenanthrene and 2-ethylanthraquinone were selected to have their K_d values measured in multiple HEMWat systems to investigate the physical factors that could affect the experimental K_d values (Figure 13). They were selected due to their solid state, high solubility in HEMWat and the fact that they have a measurable K_d values across the HEMWat systems chosen to give a broad range of the polarity. These were HEMWat 14, 17, 20 and 22 (Table 1).

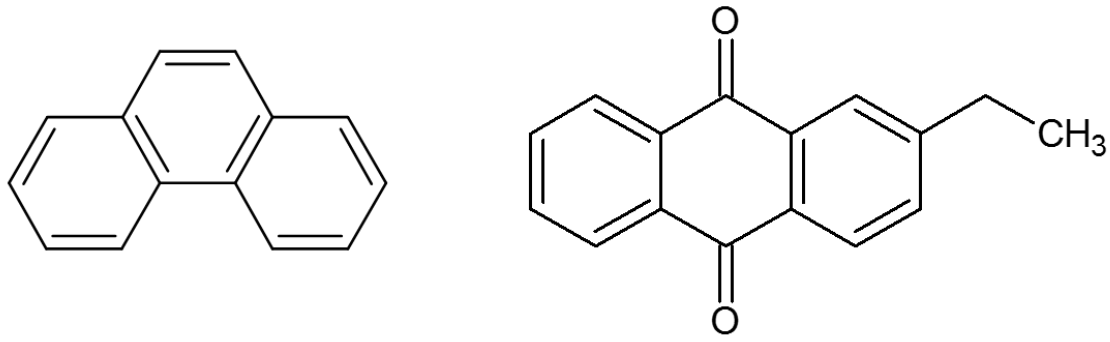


Figure 13 – The molecular structure of phenanthrene (left) and 2-ethylanthraquinone (right)

The factors investigated to ascertain the size of the impact on the experimental measurement of partition coefficient values were the addition of co-solvent, temperature, equilibrium time and pH.

All the K_d values is this section have been converted to $\log K_d$ values due to the linear relationship between $\log K_d$ and HEMWat number. This linear relationship will be exploited to link the six QSAR models together creating a model that can be used to predict the K_d values of a compound across the HEMWat series (section 5.1).

3.1. The Effect of the Addition of Co-solvent

The partition coefficient (K_d) value is determined by dividing the concentration of the compound in the upper phase divided by its concentration in the lower phase. The concentrations are taken as the peak area from the HPLC chromatogram for each phase. To ensure the accuracy of the partition coefficient values obtained, the area of the peak must have a signal-to-noise ratio greater than five. To increase the likelihood of a peak being large enough to accurately integrate, the maximum amount of compound is required in the solvent system. As many compounds have a high solubility in dimethyl sulfoxide (DMSO), this was used to initially ensure a large quantity of compound was in solution, when its K_d value was measured experimentally. However, it is known that the presence of DMSO reduces K_d values in octanol/water systems with a larger impact observed on the K_d values of hydrophobic molecules (Colclough, et al., 2015). Therefore, the level at which the presence of DMSO effects the experimental K_d value was investigated.

3.1.1. Method

Following the experimental procedure in section 2.1.2, the K_d values were experimentally determined with and without the co-solvent. In the first instance, the compounds were initially dissolved in DMSO and the K_d values experimentally measured. This was then repeated but with the initial dissolution being carried out using one phase of the HEMWat system being investigated. To maximise the amount of compound in solution, the ClogP value (predicted octanol/water partition coefficient from Daylight, Aliso Viejo, CA, USA/Biobyte, Claremont, CA, USA) of the compound was used to decide which phase of the HEMWat system was used to dissolve it. A negative ClogP meant that the compounds were dissolved in the lower phase and with compounds dissolved in the upper phase if they have a positive ClogP. As phenanthrene and 2-ethylanthraquinone have positive ClogP values, the phase used to carry out this initial dissolution was the upper phase.

3.1.2. Results and Conclusion

As can be seen in Table 4, the difference between the K_d values for 2-ethylanthraquinone obtained when a co-solvent was added and when it was not, rises as the HEMWat number increases, with the smallest percentage difference being 1% for HEMWat 14 and the largest being 48% for HEMWat 22 (Table 4). The K_d values of phenanthrene are more affected by the presence of DMSO than the values of 2-ethylanthraquinone, as there is a larger difference between the K_d values. This is as expected due it the hydrophobic nature of phenanthrene (Colclough, et al., 2015). The smallest percentage difference for phenanthrene is 26% for HEMWat 17 with the largest being 78% for HEMWat 22 (Table 5). There appears to be a general trend of the co-solvent having a larger impact on the experimental logK_d values in more non-polar systems. As accuracy and reproducibility are so important to building an accurate QSAR model, it was therefore decided to carry out all further K_d determinations without the presence of a co-solvent.

Table 4 – The experimentally determined logK_d values for 2-ethylanthraquinone in four HEMWat systems when initially dissolved in DMSO or in the upper phase of the HEMWat system.

HEMWat System Number	logK_d values when initially dissolved in DMSO	logK_d values when initially dissolved in Upper Phase HEMWat	Difference (Δ)	Percentage difference (%)
14	1.91	1.92	0.01	1
17	1.07	1.18	0.11	9
20	0.60	0.71	0.11	15
22	0.23	0.34	0.11	48

Table 5 – The experimentally determined logK_d values for phenanthrene in four HEMWat systems when initially dissolved in DMSO or in the upper phase of the HEMWat system.

HEMWat System Number	logK_d values when initially dissolved in DMSO	logK_d values when initially dissolved in Upper Phase HEMWat	Difference (Δ)	Percentage difference (%)
14	1.32	1.88	0.56	30
17	0.68	0.92	0.24	26
20	0.05	0.32	0.27	84
22	-0.40	-0.09	0.31	78

3.2. The Effect of Temperature

It is well known that any change in temperature affects solvents' density and, as a result, their mutual solubility/miscibility. In the case of HEMWat systems, methanol is soluble in water, therefore, temperature variation will lead to a change in the composition of a HEMWat system. For this reason, the sensitivity of the experimental K_d values to changes in temperature was assessed.

3.2.1. Method

Following the experimental method details in section 2.1.3, the partition coefficient values for 2-ethylanthraquinone and phenanthrene were measured at 20°C, 25°C and 30°C. This range was chosen as it is the most likely temperature variation found in a laboratory.

3.2.2. Results and Conclusion

It is demonstrated in Figure 14 and Table 6 that for phenanthrene there is very little effect on $\log K_d$ between 25°C and 30°C, as the $\log K_d$ values for all four HEMWat systems differ by less than 0.05.

Table 6 - The average and standard deviation of the $\log K_d$ values obtained at 20°C, 25°C and 30°C for phenanthrene

HEMWat Number	Temperature	Average of triplicate $\log K_d$ value	Standard Deviation of triplicate $\log K_d$ values
14	20	2.15	0.08
	25	1.90	0.02
	30	1.86	0.03
17	20	1.36	0.03
	25	0.92	0.01
	30	0.87	0.02
20	20	0.96	0.01
	25	0.34	0.01
	30	0.33	0.01
22	20	0.80	0.01
	25	-0.10	0.01
	30	-0.09	0.01

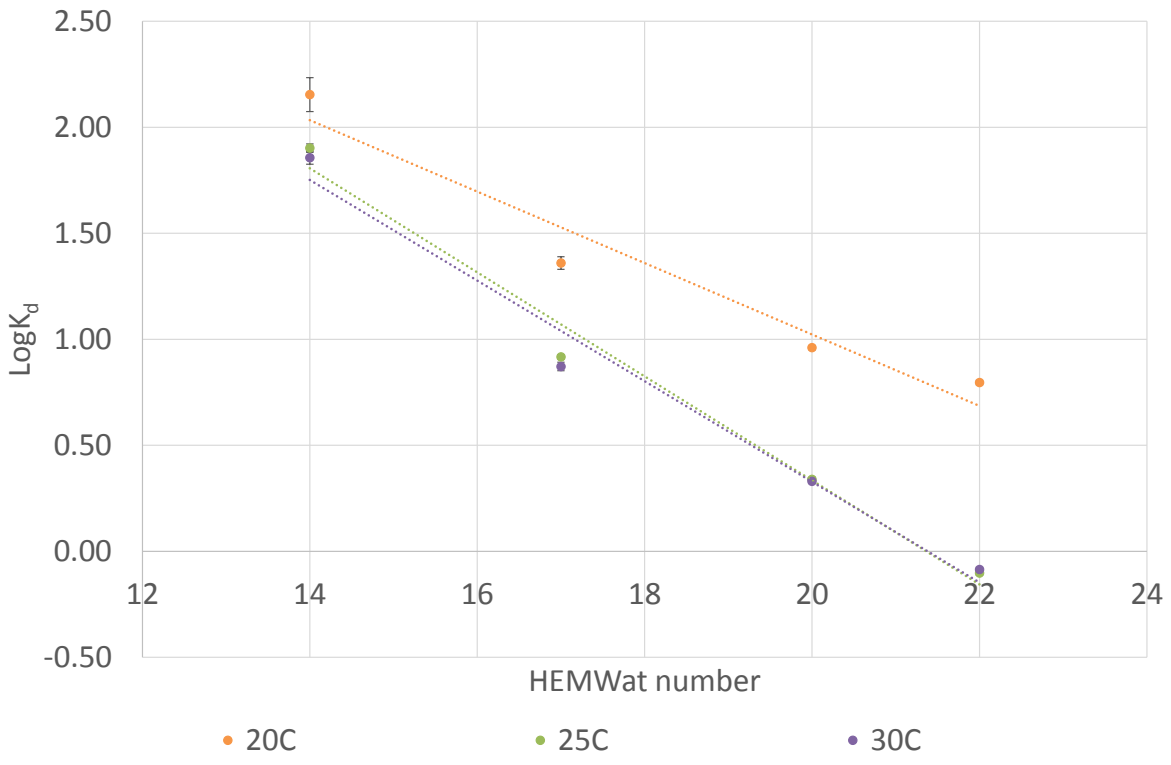


Figure 14 - The average of three logK_d values of phenanthrene at 20°C, 25°C and 30°C in HEMWat 14, 17, 20 and 22 with standard deviation error bars.

However, the drop in temperature to 20°C led to a large difference in logK_d values when compared to those obtained at 25°C and 30°C. This was particularly noticeable for more non-polar systems with the %RSD increasing from 20.97% to 206.93% between HEMWat 17 and HEMWat 22 (Table 7).

Table 7 - The average, standard deviation and %RSD of the logK_d values obtained at 20°C, 25°C and 30°C for phenanthrene

HEMWat Number	Average of triplicate logK_d value at three temperatures	Standard Deviation of triplicate logK_d values at three temperatures	%RSD of triplicate logK_d values at three temperatures
14	1.97	0.13	6.66
17	1.05	0.22	20.97
20	0.54	0.30	54.32
22	0.20	0.42	206.93

When the difference between the logK_d values for all three temperatures were compared, they were found to be smaller for 2-ethylanthraquinone (Table 8 and Figure

15) than those seen for phenanthrene (Table 7). The standard deviations for all the K_d measurements for 2-ethylanthrquinone were less than 0.03, whereas the standard deviations for phenanthrene ranged from 0.13 to 0.42.

Table 8 - The average and standard deviation of the logK_d values obtained at 20°C, 25°C and 30°C for 2-ethylanthraquinone

HEMWat Number	Temperature	Average of triplicate logK_d value	Standard Deviation of triplicate logK_d values
14	20	2.09	0.01
	25	2.14	0.03
	30	2.04	0.03
17	20	1.18	0.01
	25	1.22	0.01
	30	1.06	0.02
20	20	0.72	0.01
	25	0.71	0.01
	30	0.69	0.01
22	20	0.55	0.01
	25	0.37	0.02
	30	0.35	0.01

This is further demonstrated by the %RSD between the logK_d values for 2-ethylanthraquinone being less than 10% for HEMWat 14, 17 and 20 (Table 9). Interestingly, it is again HEMWat 22 that has the largest %RSD for the three temperatures (Figure 15).

Table 9 – The average, standard deviation and %RSD of the logK_d values obtained at 20°C, 25°C and 30°C for 2-ethylanthraquinone.

HEMWat Number	Average of triplicate logK_d value at three temperatures	Standard Deviation of triplicate logK_d values at three temperatures	%RSD of triplicate logK_d values at three temperatures
14	2.09	0.04	2.07
17	1.15	0.07	6.04
20	0.71	0.01	1.73
22	0.42	0.09	20.74

However, the logK_d values of both compounds have the same general trend with the experimentally determined logK_d values decreasing as the temperature rises.

Despite the large difference in the logK_d values obtained at 20°C and 25°C, the practical problems associated with determining a logK_d value whilst maintaining the solvent system temperature between 25°C and 30°C are considerable. As the laboratory temperature is generally maintained at 20°C, the system will be subject to a temperature change every time it is manipulated or sampled which may lead to variation in measurements. An additional consideration is the increase in evaporation of the more volatile components of the HEMWat systems leading to a change in solvent system composition. Therefore, to reduce the likelihood of variation in experimental logK_d measurements caused by a change in temperature, it was decided to conduct all future experiments at 20°C as this was likely to be closest to the room temperature. Due to the large difference between the logK_d values of phenanthrene at 20°C and 25°C, if the laboratory temperature was raised above 20°C, no experimental logK_d temperature measurements would be taken as this was likely to have a large impact on logK_d.

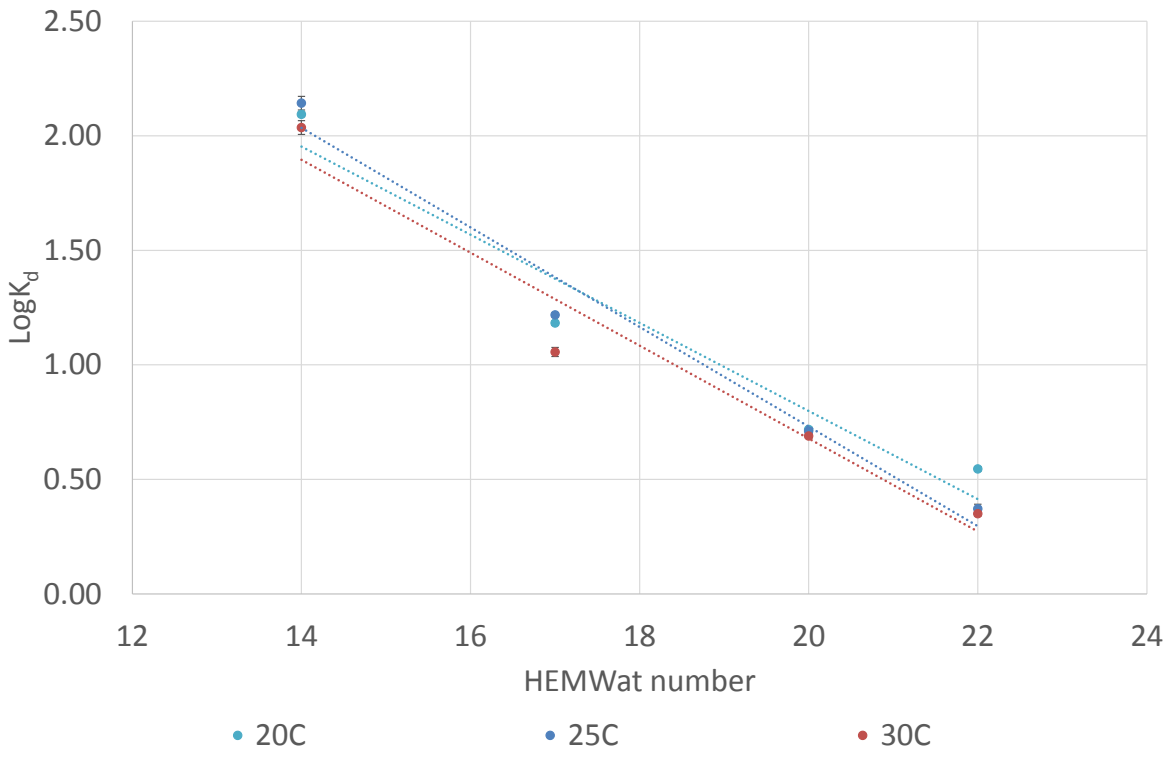


Figure 15 – The average of three logK_d values of 2-ethylanthraquinone at 20°C, 25°C and 30°C in HEMWat 14, 17, 20 and 22.

As temperature was seen to have such a large effect on the logK_d values of phenanthrene, it was decided to prepare all of the HEMWat systems by mass instead of the more conventional route of by volume. This had the added benefit of the balance allowing the measurements of the mass of each solvent to four decimal places compared to the pipettes used to measure volume which was only capable of dispensing integer volumes.

3.3. The Effect of Equilibration Time

The HEMWat systems being tested are made up of four solvents. Initially, when the solvents are mixed, the interface between them can be disrupted and an emulsion formed. However, when allowed to settle and reach equilibrium, two phases will be formed. The failure to reach equilibrium conditions can lead to a lack of mutual phase saturation, which will add error to the experimental K_d value. If the solvent system has not reached equilibrium, the K_d measurement may be taken when the phases are not of constant composition, therefore altering the K_d value. This would lead to inaccuracy and poor reproducibility.

3.3.1. Method

Following the experimental method details in section 2.1.4, the partition coefficient values for 2-ethylanthraquinone and phenanthrene were measured in HEMWat 17 (Table 1) as this system is a common starting point for CCC method development. The system was vigorously mixed then left to equilibrate and form two phases. At five different time points, the upper and lower phases were sampled and partition coefficient measurements were taken.

3.3.2. Results and Conclusion

As can be seen in Table 10, the five $\log K_d$ values measured between 0 minutes and overnight for 2-ethylanthraquinone and phenanthrene are very consistent with a %RSD of 3.72% for 2-ethylanthrquinone and a %RSD for phenanthrene of 6.30%, both of which are within the acceptance criteria of 10%. This suggests that the amount of time that the HEMWat systems are left to reach equilibrium does not affect the $\log K_d$ value of 2-ethylanthraquinone and phenanthrene.

Table 10 – The partition coefficient values of 2-ethylanthraquinone and phenanthrene in HEMWat 17 that had been allowed to settle for different periods of time. The first measurement was taken as soon as two phases had formed after mixing (0 minutes). The remaining four time points were the length of time the systems had been left to equilibrate and were measured from when the two phases first formed.

Time elapsed after mixing was stopped	Average $\log K_d$ values of triplicate measurements	
	2-Ethylanthraquinone	Phenanthrene
0 minutes	1.06	1.22
30 minutes	1.14	1.34
1 hour	1.17	1.40
2 hours	1.15	1.19
Overnight	1.18	1.36

However, the solvent system remained cloudy and did not become fully transparent until it had been left to equilibrate overnight. This suggests that full equilibrium was not reached. Despite there being no change observed in the partition coefficient values of these two compounds in HEMWat 17, other compounds in other solvent systems may be more greatly affected by this. As the accuracy of a QSAR model is so dependent

on the accuracy of the K_d measurements, it was decided that all the solvent systems should be left to equilibrate overnight before use.

3.4. The Effect of Solute Concentration

To ensure the accuracy of the K_d value measurements, the compounds were dissolved in one of the phases of a solvent system to the point of saturation before mixing it with another phase. This aimed to obtain HPLC peaks large enough to integrate and minimise the error. However, it had been suggested by Dearden and Bresnan that when measuring partition coefficient values, the concentration of a solute should be less than 10^{-3} M and for carboxylic acids it should be less than 10^{-4} M for non-polar phases (Dearden, et al., 1988). This was suggested to prevent solute self-association whilst maintaining constant phase composition and allowing the assumption that activity coefficients are close to unity (Dearden, et al., 1988). However, Kenny et al. stated that accurate K_d measurements could be obtained using saturated solutions as long as the compounds do not have a tendency to dimerise (Kenny, et al., 2013), for example, carboxylic acids.

3.4.1. Method

Following the method detailed in section 2.1.5, phenol, warfarin and 3-bromobenzoic acid were selected to carry out this assessment (Figure 16). These three compounds were chosen as they are structurally diverse and include a carboxylic acid, a compound class known to dimerise at high concentrations (Dearden, et al., 2009).

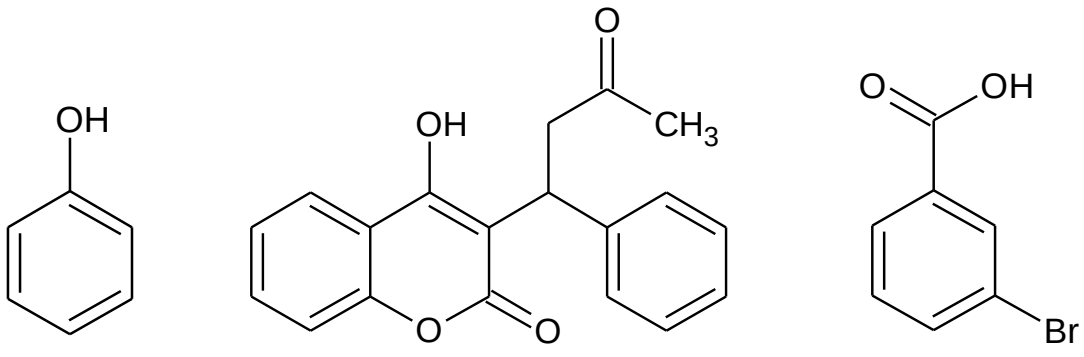


Figure 16 – The molecular structure of phenol (left), warfarin (middle) and 3-bromobenzoic acid (right)

The phase used to carry out the initial dissolution was saturated with each compound and the K_d value was measured. This phase was diluted four times with a fresh portion of the phase and for each dilution the partition coefficient values were measured.

3.4.2. Results and Conclusion

As can be seen from Table 11, altering the concentration only leads to a small change in the partition coefficient values for phenol and 3-bromobenzoic acid, with the compounds having a %RSD across the concentration range of 2.51% and 1.18% respectively. The variation is higher for warfarin with a %RSD across the concentration range of 24.94%. The standard deviations are below 0.1 for all concentrations except one, suggesting good reproducibility from the experimental measurements. However, as there is no constant pattern, it was therefore concluded that the concentrations used in this method did not affect the K_d value and the K_d measurements using the initial saturation of one phase was continued.

Table 11 – The partition coefficient values of phenol, warfarin and 3-bromobenzoic acid measured using five different initial concentrations in HEMWat 17 prepared by mass and left to equilibrate overnight.

Compounds	Concentration (M)	Average K_d value from triplicate measurements	Standard Deviation K_d value	%RSD K_d value
Phenol	0.531	0.66	0.04	6.40
	0.266	0.62	0.01	1.71
	0.133	0.65	0.02	2.82
	0.066	0.62	0.03	4.43
	0.033	0.64	0.01	1.48
Warfarin	0.08	0.59	0.04	6.26
	0.04	0.48	0.02	4.07
	0.02	0.49	0.02	3.88
	0.01	0.88	0.17	19.17
	0.005	0.53	0.02	2.94
3-Bromobenzoic acid	0.24	1.12	0.08	6.91
	0.12	1.15	0.02	1.70
	0.06	1.13	0.02	1.80
	0.03	1.13	0.01	1.02
	0.015	1.11	0.03	2.51

Having established that the concentration of the solution does not impact the K_d value when measured by HPLC, this was confirmed by measuring K_d values by CCC. To

follow the same dilution pattern, four sample concentrations were investigated, starting from the saturated solution of a compound followed by three one-to-one dilutions. As can be seen in Figure 17, the elution times of the 3-bromobenzoic acid vary by 0.2 minutes. This change in elution time is not large enough to change the K_d value obtained from the CCC chromatogram. Therefore, the partition coefficient value is independent of the initial concentration of the injected sample.

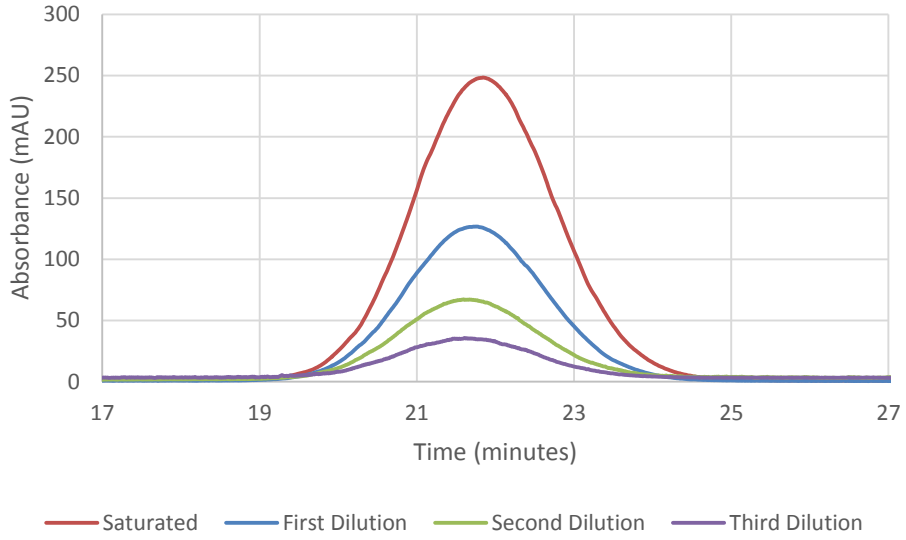


Figure 17 – The elution profile of 3-bromobenzoic acid at four different concentrations in HEMWat 17 from a saturated solution through three stepwise 1:1 dilutions with fresh lower phase. The concentration of the saturated solution was 0.24M, with the concentration of the first, second and third dilution being 0.12, 0.06 and 0.03M. The range of the four elution times is 0.2 minutes.

3.5. Conclusion

It was concluded that to try to ensure the most accurate K_d measurements no co-solvent would be used, the measurements would be taken at 20°C, with the solvent systems left overnight to equilibrate. As the concentration of the compound was not found to influence the K_d measurement, they were taken at saturation. In addition, it was decided to extend the range of HEMWat systems being examined to include HEMWat 8 and 26 to increase the spread of system polarity being investigated.

3.6. Reproducibility of K_d measurements

The reproducibility of the experimentally determined K_d values was assessed using the developed methodology (see section 3.7.5) whilst making possible changes in

equipment set up or solvents used. This would ensure that the developed procedure can be transferred to any laboratory conditions.

Three compounds had their K_d values determined in triplicate at 20°C, in solvent systems equilibrated overnight. The measurements of the K_d values for the same compounds were then repeated using the standard experimental methodology. The differences were:

- New batch of solvents
- Different laboratory
- New HPLC column
- New HPLC needle
- New HPLC needle seat
- Minimum of 3 months apart

The compounds tested were chosen on the basis of their structural diversity. Caffeine was chosen due to its highly functionalised nature, phenanthrene as it has no functionality but is aromatic and 2-ethylanthraquinone as it has certain functionality and aromaticity (Figure 18).

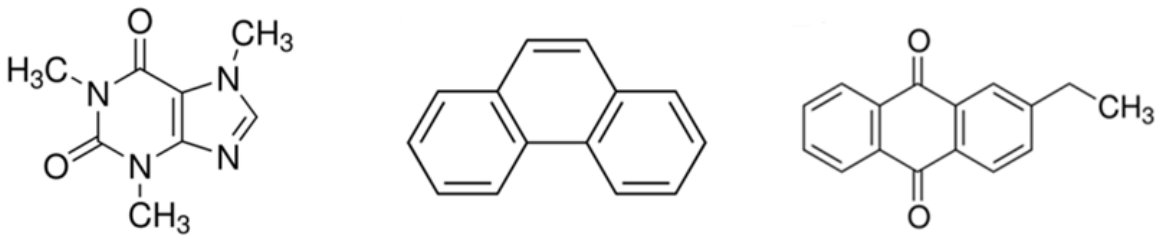


Figure 18 - The molecular structures of caffeine (left), phenanthrene (middle) and 2-ethylanthraquinone (right).

It was found that all experimental K_d values were reproduced satisfactorily with low %RSD of less than 10% values for all measurements for caffeine and phenanthrene (Table 12). In the case of 2-ethylanthraquinone, four out of the six solvent systems had %RSD values of less than 10%. The K_d values for 2-ethylanthraquinone in HEMWat 20 and gave a %RSD value of 11.98% and 14.04%. However, the standard

deviation is less than 0.1 in all cases so the reproducibility of the K_d measurements was deemed acceptable.

Table 12 – The average, standard deviation and relative standard deviation of six experimentally determined logK_d values of three neutral compounds, three initial and three repeats in six HEMWat systems.

HEMWat system number	2-Ethylanthraquinone			Phenanthrene			Caffeine		
	Average	Standard Deviation	%RSD	Average	Standard Deviation	%RSD	Average	Standard Deviation	%RSD
8	3.49	0.08	2.43	3.24	0.07	2.03	-0.40	0.03	7.90
14	2.15	0.05	2.26	2.15	0.08	3.79	-0.94	0.02	1.64
17	1.18	0.06	5.33	1.36	0.03	1.98	-1.49	0.02	1.33
20	0.71	0.08	11.98	0.96	0.01	0.98	-1.90	0.02	1.27
22	0.52	0.04	7.33	0.80	0.01	0.93	-1.99	0.02	0.85
26	0.20	0.03	14.04	0.42	0.01	1.33	-1.92	0.04	1.86

To further investigate the robustness of the developed methodology, the same approach as above was applied to a set of three acidic compounds. These were ibuprofen, warfarin and tolbutamide (Figure 19) which were chosen for their structural diversity with ibuprofen containing a carboxylic acid group, warfarin containing an alcohol group and tolbutamide with a carbonyl group directly connected to two amine groups. The pKa values for the three compounds were 4.91 for ibuprofen (Drugbank, 2015), 5.08 for warfarin (Drugbank, 2015) and 5.16 for tolbutamide (Drugbank, 2015).

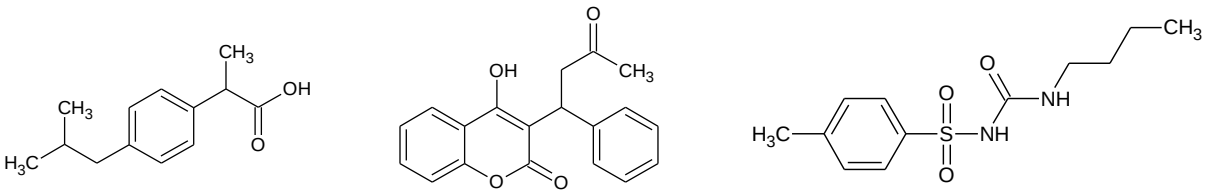


Figure 19 - The molecular structure of ibuprofen (left), warfarin (middle) and tolbutamide (right)

The %RSD for all six of these measurements can be found in Table 13. The majority of the experimental K_d measurements reproduced well with a %RSD of less than 10% for ibuprofen and tolbutamide. However, the logK_d values for ibuprofen in HEMWat 20

and for tolbutamide in HEMWat 14 and 17 had RSD% values of 25.33%, 12.66% and 11.21% respectively. In both cases, the HEMWat system, in which the variation occurs, is the system in which the $\log K_d$ is close to 0. Despite the standard deviation across all six systems being less than 0.1, the %RSD is much larger for the systems in which the compound has a $\log K_d$ value of zero. As stated in the literature review, separation is most likely to be achieved using CCC when the K_d value of a compound is between 0.5 and 2 (Ren, et al., 2013). The value of K_d of 1 was selected for prediction as it is within this range. However, a system in which the compound has a K_d value of 1 has a $\log K_d$ of 0, therefore, despite all the systems having standard deviations below 0.1, the %RSD will be higher when $\log K_d$ is 0. As this is the value being predicted by the models, it is important that this measurement is very accurate.

Table 13 - The average, standard deviation and relative standard deviation of six experimentally determined $\log K_d$ values of tolbutamide, ibuprofen and warfarin in six HEMWat systems. There is no $\log K_d$ value for ibuprofen in HEMWat 8 as these values were so extreme one of the peaks present in the HPLC chromatogram did not have a signal-to-noise ratio greater than five. This meant it was deemed too small to integrate and a partition coefficient value was not obtained in this system.

HEMWat system number	Tolbutamide $\log K_d$ values			Ibuprofen $\log K_d$ values			Warfarin $\log K_d$ values		
	Average	Standard Deviation	%RSD	Average	Standard Deviation	%RSD	Average	Standard Deviation	%RSD
8	2.17	0.01	0.58	-	-	-	2.50	0.26	10.43
14	0.59	0.07	12.66	1.69	0.04	2.39	1.04	0.32	30.76
17	-0.62	0.07	11.21	0.63	0.02	3.02	-0.89	0.47	52.43
20	-1.56	0.09	5.84	-0.08	0.02	25.33	-1.63	0.43	26.42
22	-1.93	0.03	1.64	-0.28	0.03	9.07	-1.81	0.27	14.83
26	-2.02	0.05	2.36	-0.63	0.03	4.77	-2.20	0.32	14.76

It was hypothesised that the problem of the lack of reproducibility of K_d values around one could be caused by a difference in partitioning between the neutral and acidic forms of the ionisable molecules and the sensitivity of the proportion of the neutral form to pH. Adding acid to the HEMWat system could potentially solve this problem, as the acid would force the molecules all into the neutral state. To keep the volume of the additional fifth component in the system to a minimum (section 3.1), 0.1% v/v TFA

in water was used to prepare the HEMWat systems, instead of pure water. The original three experimentally determined $\log K_d$ were compared to three new $\log K_d$ values generated using acidified HEMWat. From Table 14, Table 15 and Table 16, it can be seen that by adding the TFA into the water, the variation in results was reduced in all but one of the cases. Most encouragingly, the %RSD values for the systems in which the three compounds have a $\log K_d$ value close to zero were reduced. However, for warfarin in HEMWat 17 and for ibuprofen in HEMWat 20 the %RSD value were still above 10% (29.90% and 12.37%). In the cases where the addition of acid had an adverse effect of the variation, none of the percentage differences were as large as the reduction seen in ibuprofen and warfarin.

Table 14 - The average, standard deviation and relative standard deviation of the partition coefficient values of tolbutamide in unadjusted HEMWat and acidified HEMWat (0.1%TFA in water replacing water) in six HEMWat systems.

HEMWat system number	Tolbutamide $\log K_d$ values in triplicate					
	Unadjusted HEMWat			HEMWat with 0.1% TFA in the water		
	Average	Standard Deviation	%RSD	Average	Standard Deviation	%RSD
8	2.17	0.01	0.58	2.24	0.00	0.08
14	0.59	0.07	12.66	0.79	0.01	0.87
17	-0.62	0.07	11.21	-0.44	0.01	2.11
20	-1.56	0.09	5.84	-1.30	0.02	1.74
22	-1.93	0.03	1.64	-1.71	0.02	1.11
26	-2.02	0.05	2.36	-1.84	0.00	0.16

Table 15 - The average, standard deviation and relative standard deviation of the partition coefficient values of ibuprofen in unadjusted HEMWat and acidified HEMWat (0.1%TFA in water replacing water) in six HEMWat systems. There is no logK_d value for ibuprofen in HEMWat 8 as these values were so extreme one of the peaks present in the HPLC chromatogram did not have a signal-to-noise ratio greater than five. This meant it was deemed too small to integrate and a partition coefficient value was not obtained in this system.

HEMWat system number	Ibuprofen logK_d values in triplicate					
	Unadjusted HEMWat			HEMWat with 0.1% TFA in the water		
	Average	Standard Deviation	%RSD	Average	Standard Deviation	%RSD
8	-	-	-	-	-	-
14	1.66	0.05	2.90	1.83	0.02	0.94
17	0.67	0.05	7.85	0.62	0.01	1.19
20	0.16	0.20	123.40	-0.04	0.01	29.90
22	-0.28	0.02	7.00	-0.28	0.02	8.63
26	-0.63	0.04	6.06	-0.61	0.02	3.12

Table 16 - The average and standard deviation of the partition coefficient values of warfarin in unadjusted HEMWat and acidified HEMWat (0.1%TFA in water replacing water) in six HEMWat systems.

HEMWat system number	Warfarin logK_d values in triplicate					
	Unadjusted HEMWat			HEMWat with 0.1% TFA in the water		
	Average	Standard Deviation	%RSD	Average	Standard Deviation	%RSD
8	2.50	0.26	10.43	2.82	0.04	1.48
14	1.04	0.32	30.76	0.90	0.04	4.05
17	-0.89	0.47	52.43	-0.85	0.11	12.37
20	-1.63	0.43	26.42	-1.42	0.01	0.75
22	-1.81	0.27	14.83	-1.66	0.08	4.91
26	-2.20	0.32	14.76	-1.96	0.04	2.18

It was considered that the same problem may be affecting bases. Therefore, the water in the HEMWat system was replaced with 1% v/v ammonia solution (33% v/v) to form basified HEMWat. This percentage was selected by the fact that 0.1% TFA is pH 4, which is three pH units away from pH 7 and, therefore, ammonia was added to the water until it reached pH 10. The $\log K_d$ values for three bases in basified HEMWat were compared to the experimental $\log K_d$ values in unadjusted HEMWat. The three bases chosen were lidocaine, nadolol and reserpine as shown in Figure 8. Reserpine was selected as it is one of the larger molecules in the training set, with lidocaine and nadolol chosen due their differing number of nitrogen atoms. The pKa values for reserpine, lidocaine and nadolol are 6.6 (Drugbank, 2015), 8.01 (Drugbank, 2015) and 9.67 (Drugbank, 2015) respectively.

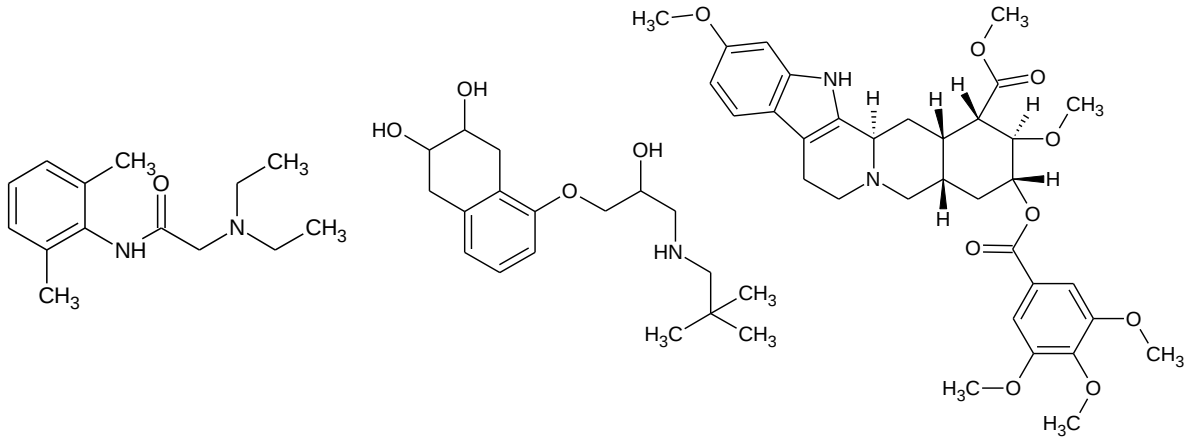


Figure 20 – The molecular structure of lidocaine (left), nadolol (middle) and reserpine (right).

The variation in $\log K_d$ values did not seem to have as great an impact on bases but for the $\log K_d$ value of lidocaine in HEMWat 8, the %RSD was reduced from 25.40 to 0.95% with the addition of 1% NH_4OH solution to the water (Table 17). However, running the lidocaine compounds in basified HEMWat only reduced the variation in a two out of the six systems. Despite this, for five out of the six systems the %RSD in basified HEMWat were below 10%. The standard deviations were below 0.1 is all cases.

Table 17 – The average, standard deviation and relative standard deviation (%RSD) of lidocaine in unadjusted HEMWat and basified HEMWat (1% v/v NH_4OH solution (33% v/v) in water replacing water) in six HEMWat systems.

HEMWat system number	Lidocaine $logK_d$ values in triplicate					
	Unadjusted HEMWat			Basic HEMWat		
	Average	Standard Deviation	%RSD	Average	Standard Deviation	%RSD
8	0.10	0.03	25.40	1.98	0.02	0.94
14	-0.66	0.01	1.44	1.07	0.00	0.31
17	-1.46	0.02	1.12	0.09	0.01	11.11
20	-2.10	0.01	0.39	-0.50	0.02	4.40
22	-2.24	0.02	0.92	-0.65	0.03	4.92
26	-2.27	0.01	0.55	-0.88	0.01	1.44

For nadolol, the %RSD for the $logK_d$ values in two of the three systems has increased. However, in both these cases the standard deviation has been reduced or maintained (Table 18).

Table 18 – The average, standard deviation and relative standard deviation (%RSD) of nadolol in unadjusted HEMWat and basified HEMWat (1%v/v NH_4OH solution (33% v/v) in water replacing water) in six HEMWat systems. There are no $logK_d$ values for nadolol in HEMWat 20, 22 and 26 as these values were so extreme one of the peaks present in the HPLC chromatogram did not have a signal-to-noise ratio greater than five. This meant it was deemed too small to integrate and a partition coefficient value was not obtained in these systems.

HEMWat system number	Nadolol $logK_d$ values in triplicate					
	Unadjusted HEMWat			Basic HEMWat		
	Average	Standard Deviation	%RSD	Average	Standard Deviation	%RSD
8	-1.56	0.04	2.41	0.07	0.00	5.74
14	-1.76	0.02	1.02	-0.20	0.02	10.11
17	-2.30	0.03	1.37	-1.12	0.00	0.00
20	-	-	-	-	-	-
22	-	-	-	-	-	-
26	-	-	-	-	-	-

For reserpine, two of the systems show an increase in the %RSD values. However, in one case the %RSD is still below 10% and in the other the standard deviation has been reduced.

Table 19 – The average, standard deviation and relative standard deviation (%RSD) of reserpine in unadjusted HEMWat and basified HEMWat (1%v/v NH_4OH solution (33% v/v) in water replacing water) in six HEMWat systems. There are no $logK_d$ values for reserpine in HEMWat 20, 22 and 26 as these values were so extreme one of the peaks present in the HPLC chromatogram did not have a signal-to-noise ratio greater than five. This meant it was deemed too small to integrate and a partition coefficient value was not obtained in these systems.

HEMWat system number	Reserpine $logK_d$ values in triplicate					
	Unadjusted HEMWat			Basic HEMWat		
	Average	Standard Deviation	%RSD	Unadjusted HEMWat	Basic HEMWat	%RSD
8	2.81	0.08	2.75	2.46	0.04	1.76
14	1.66	0.03	1.98	1.49	0.05	3.04
17	-0.37	0.05	12.65	-0.10	0.03	25.71
20	-	-	-	-	-	-
22	-	-	-	-	-	-
26	-	-	-	-	-	-

Although these results do not show the large reduction in variation seen when the acids were run in acidified HEMWat, it was decided to run the bases in basified HEMWat. This was because the pH of a HEMWat systems may vary from laboratory to laboratory which may affect the experimental measurement of the $logK_d$ values of bases. By adding a pH modifier and making the pH of the HEMWat systems more extreme, it is likely that compounds that are sensitive to pH changes, will be less affected by laboratory to laboratory variation.

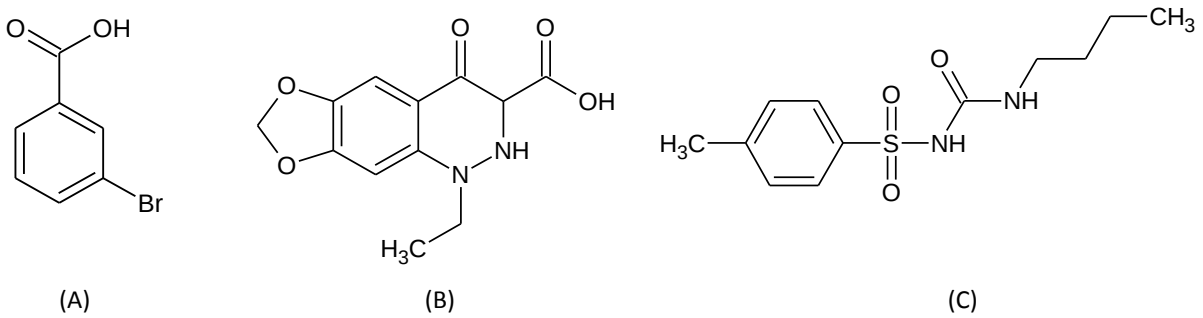
3.7. The Effect of pH

As it had been decided to use acidified and basified HEMWat for measuring the experimental partition coefficient values of acidic and basic compounds, the effect of pH on the partition coefficient was examined in more detail. It had been noted in the literature by Dearden and Bresnen that a change in pH could have a large impact on

partition coefficient values (Dearden, et al., 1988). In addition Conway stated that, for a unit of pH change, there could be a 10 fold change in K_d value (Conway, 1991). The logD of a compound takes into account the neutral and ionised species of a compound (section 1.8.4.1.1). However, the model will only predict logK_d values for the neutral species due to the addition of pH modifier. The level of the pH modifiers used in the section 3.5 were kept low, with 0.1% v/v TFA or 1% v/v NH_4OH (33% v/v) in water, replacing the water in the HEMWat systems. This low level was specifically chosen to minimise the effect of the presence of this additional component in the systems (see section 3.1), whilst neutralising acidic or basic compounds. It had been hypothesised that the decrease in variation in the experimental logK_d measurement, was due to all the molecules being neutralised. To confirm this, HEMWat systems were made up with the water containing three concentrations of TFA: 0.1%, 0.4% and 0.8%, by volume. If the 0.1% TFA in water contained enough acid to have fully neutralised the molecules, there would be no difference in K_d values observed. Please refer to section 2.1.6 for details on how the experimentally determined K_d values were obtained.

3.7.1. Investigating the impact of the addition of different amounts of TFA to the water of the HEMWat systems

The partition coefficient values of 3-bromobenzoic acid, cinoxacin and tolbutamide (Figure 21) were measured on the CCC using three concentrations of TFA added to the water of the selected HEMWat systems. The three concentrations were 0.1%, 0.4% and 0.8% measured by volume. All three compounds were chosen due to their low K_d values (less than 2) in HEMWat 17 (see section 1.3). Tolbutamide had the added benefit of being in the original three compounds used to investigate the effect of the addition of TFA to a HEMWat system. Cinoxacin and 3-bromobenzoic acid also met the criteria of a K_d value of less than 2 in HEMWat 17 and are both carboxylic acids which had been identified in the literature as being particularly affected by pH changes. The pKa values of tolbutamide, cinoxacin and 3-bromobenzoic acid are 5.16 (Drugbank, 2015), 4.6 (Koike, et al., 1984) and 3.81 (Chemicalbook, 2015) respectively.



(A) (B) (C)

Figure 21 – The molecular structure of 3-bromobenzoic acid (A), cinoxacin (B) and tolbutamide (C).

The HEMWat systems were prepared by mass and left to equilibrate overnight. The experimental method details can be found in section 2.3.2. Table 20 shows that there was no change in the K_d values of cinoxacin. However, there was a %RSD of 33% in the K_d values for 3-bromobenzoic acid. Although a similar change was observed for tolbutamide (35%), in practical terms, a change in K_d values between 0.3 and 0.2 is not generally taken into account. This decrease in K_d values was unexpected as a greater proportion of neutralised molecules of the compound the K_d values would increase. This may be due to the level of TFA being too low to prevent variation.

Table 20 – The partition coefficients of 3-bromobenzoic acid, cinoxacin and tolbutamide measured using CCC with HEMWat 17 solvent system, which was prepared by mass with 0.1% TFA added to the water only and was left overnight to equilibrate in a separating funnel.

Compounds	Partition coefficient values		
	0.1% TFA by volume added to the water used to make up the HEMWat system	0.4% TFA by volume added to the water used to make up the HEMWat system	0.8% TFA by volume added to the water used to make up the HEMWat system
3-Bromobenzoic acid	1.2	1.0	0.8
Cinoxacin	0.0	0.0	0.0
Tolbutamide	0.3	0.3	0.2

3.7.2. Investigating the impact of the addition of different amounts of TFA into all four solvents of the HEMWat system

The amount of acid in each HEMWat system was increased by adding TFA into all four solvents to see if there was an impact on K_d value as opposed to it being added to the water alone. This had the added benefit of ensuring that each HEMWat system contained the same amount of acid, as when the acid is added to the water only,

systems with lower water content will contain less acid. It was found that the K_d values of 3-bromobenzoic acid, cinoxacin and tolbutamide increased as the amount of acid increased (Table 21). This was expected as neutral molecules preferentially partition into the organic phase.

Table 21 – Following the experimental method details in section 2.1.6, the partition coefficient values for 3-bromobenzoic acid, cinoxacin and tolbutamide measured using CCC. The solvent system, HEMWat 17, was prepared by mass with the acid present in all four solvents before the HEMWat systems were made up. The solvent system was left to equilibrate overnight before CCC was run.

Compounds	Partition coefficient values			
	Unadjusted HEMWat	0.1% TFA by volume added to the all the solvents used to make up the HEMWat system	0.4% TFA by volume added to the all the solvents used to make up the HEMWat system	0.8% TFA by volume added to the all the solvents used to make up the HEMWat system
3-Bromobenzoic acid	0.7	0.9	1.0	1.2
Cinoxacin	0.4	0.0	0.0	0.1
Tolbutamide	0.1	0.1	0.2	0.2

However, these results presented a further complication as there was a difference in experimental K_d value between the values measured using CCC and the values measured using HPLC (Table 22).

Table 22 – The experimentally determined partition coefficient of 3-bromobenzoic acid using CCC and HPLC in HEMWat17 that had been made by mass and equilibrated with TFA in all solvents. NR indicates that the measurement could not be made due to one of the HPLC peaks not having a signal-to-noise ratio greater than five.

Compound	Analysis method	Partition coefficient values		
		HEMWat system containing 0.1% TFA v/v	HEMWat system containing 0.4% TFA v/v	HEMWat system containing 0.8% TFA v/v
3-Bromobenzoic acid	CCC	0.9	1.0	1.2
	HPLC	0.2	0.2	NR
Tolbutamide	CCC	0.1	0.2	0.2
	HPLC	0.3	0.2	0.4
Cinoxacin	CCC	0.0	0.0	0.1
	HPLC	0.0	0.0	0.0

To determine whether this was a pH effect, the K_d values of three neutral compounds were measured in HEMWat 26. The compounds were chosen for their lack of functionality to avoid any pH effect. However, this meant that they had very extreme K_d values in five out of the six HEMWat systems. Therefore, it was decided to carry out the CCC runs in HEMWat 26 as the HPLC K_d values suggested that this system would provide a reasonable run time. As can be seen in Table 23, there is a good match between the partition coefficient values determined by HPLC and those measured by CCC.

Table 23 - The partition coefficient values of neutral compounds in HEMWat 26 measured by HPLC and CCC. The solvent systems for the HPLC measurement were made up by volume and left to equilibrate overnight, whereas the solvent systems for CCC were prepared by "mixing on demand".

Compound	Analytical Method	Partition coefficient value
Biphenyl	CCC	2.5
	HPLC	2.4
Phenanthrene	CCC	2.2
	HPLC	2.6
2-Ethylanthraquinone	CCC	1.2
	HPLC	1.4

These results suggested that the discrepancy between the HPLC and CCC K_d values for 3-bromobenzoic acid was due to a pH effect, the K_d values of two more carboxylic acids were determined by HPLC and CCC and there was found to be a difference between the measured K_d values (Table 24).

Table 24 – The experimentally determined partition coefficient of naproxen and aspirin using CCC and HPLC in HEMWat17 that had been made by mass and equilibrated with TFA in all solvents.

Compound	Analysis method	Partition coefficient values		
		HEMWat system containing 0.1% TFA v/v	HEMWat system containing 0.4% TFA v/v	HEMWat system containing 0.8% TFA v/v
Naproxen	CCC	1.4	1.6	1.7
	HPLC	0.7	0.9	1.3
Aspirin	CCC	0.3	0.3	0.3
	HPLC	0.2	0.2	0.3

It was hypothesised that the change in the partition coefficient values of carboxylic acids measured by CCC could be due to the mobile phase during the run providing a constant supply of fresh acid. This class of compounds had been identified in the

literature as being particularly susceptible to pH change (Dearden, et al., 1988) (Kenny, et al., 2013). This fresh, un-dissociated acid in the CCC run will affect the equilibrium position of the acid dissociation equation (Equation 51) leading to a greater proportion of neutral molecules. There is a limited amount of acid in the HPLC via so there may be less protons available to neutralise the compound than in a CCC run.

3.7.3. Investigating the impact of the pH of both phases throughout a run

As it had been hypothesised that there may be additional acid available in the CCC column allowing a larger proportion of a compound's molecules to be neutralised, the pH of the mobile and stationary phases were measured throughout the run using a Mettler Toledo "in lab micro" pH meter. If the pH of both phases is maintained during the run, this may indicate that the proportion of the compound's molecules that are neutral is likely to be consistent throughout the experiment. If the pH increases during the run, this may suggest that a smaller proportion of molecules are neutralised at the end of the run compared to the beginning, with a decrease in pH indicating the opposite. The compounds used during this CCC run were 3-bromobenzoic acid and uracil. The 3-bromobenzoic acid was used as its K_d value had been shown to be affected increasing the TFA concentration. The uracil was added to mark the solvent front in the run to allow the pH before and after this point to be assessed.

As can be seen in Table 25, the pH is lower in the HEMWat system that was prepared with all the four solvents containing acid, which is as expected. Interestingly, the range in pH of the lower phases is 0.06 pH units whether the TFA was added to all of the solvents of the HEMWat system or just the water. It is also worth noting that the pH in the upper phase is less consistent with a range of 0.44 when the acid was added to all of the solvents of the HEMWat system and 1.35 when the acid was added to the water only. However, this is likely to be due to the high organic content of this phase. It has been stated in the literature that even trace amounts of water in organic solvents could lead to large alterations in the measurements of pH (Himmel, et al., 2011). This may have caused instability in the pH readings of the upper phase. Therefore, it would be difficult to draw a solid conclusion due to practical challenges in accurate measurement of pH in the solvent systems containing organic solvents.

Table 25 – The pH measurement of the upper and lower phase of HEMWat 17 during a CCC run using uracil and 3-bromobenzoic acid. Uracil was added as a marker for the solvent front and to check that having more than one compound, did not add to any changes. The experimental details can be found in section 2.3.3.

The point in the CCC run at which the pH was measured	pH	
	TFA all solvents 0.1% mixing on demand	TFA water only 0.1% mixing on demand
Injected sample dissolved in HEMWat lower phase (TFA in water only)	3.15	3.15
Fresh stationary upper phase (used to fill CCC column)	2.04	4.09
Displaced stationary upper phase (whilst equilibrating with mobile lower phase)	1.91	4.50
Stationary upper phase (column content after run)	1.60	3.64
Fresh mobile upper phase (did not pass through CCC)	1.68	4.99
Eluted mobile lower phase (after equilibrating)	1.72	3.18
Eluted mobile lower phase measured after uracil peak	1.68	3.24
Eluted mobile lower phase measured after 3-bromobenzoic acid peak	1.66	3.19

3.7.4. Investigating the possible impact of the hydrolysis of ethyl acetate

As no firm conclusion could be drawn from the pH measurement from within the CCC machine, confirmation that the effect was a genuine pH effect was sort. It was hypothesised that by making up the solvent systems by mass and leaving them to equilibrate overnight, the hydrolysis of the ethyl acetate to acetic acid may be altering the percentage solvent composition of the system. As this is an acid catalysed reaction, the addition of TFA may be accelerating this process. This could be the cause of the variation in the experimental measurement of K_d. To determine whether the

differences were due to a pH effect or resulting from changing the solvent system composition through the hydrolysis of ethyl acetate, phenanthrene was selected to carry out this investigation as its structure has only benzene rings. This means that there are no functional groups that can be affected by the presence of acid. Therefore, any change in the measured K_d value must be due to a change in the composition of the solvent system. Following the experimental details in section 2.1.6, the partition coefficient values of phenanthrene were determined by CCC in unadjusted HEMWat 17 and HEMWat 17 with three concentrations of TFA (0.1%, 0.4% and 0.8% v/v).

Table 26 – The experimental partition coefficient values of phenanthrene in HEMWat 17 made by volume (specifically mixing on demand) with TFA in all of the solvents at three concentrations (0.1%, 0.4% and 0.8%) and unadjusted HEMWat with no additional TFA.

Compound	Partition coefficient values			
	Unadjusted HEMWat	0.1% TFA by volume added to the all the solvents used to make up the HEMWat system	0.4% TFA by volume added to the all the solvents used to make up the HEMWat system	0.8%TFA by volume added to the all the solvents used to make up the HEMWat system
Phenanthrene	28.8	28.3	29.9	28.7

As can be seen in Table 26, the difference in K_d values for the neutral phenanthrene molecule are very small, despite a change in the acid concentration. With a %RSD of 2% over the four readings, this difference is not considered significant as it is less than 10%. Therefore, this suggests that the differences are not due to a change in the composition of the solvent system.

3.7.5. Investigating the impact of the porous nature of the Teflon tubing used within CCC machines.

An additional concern was that the porous Teflon tubing used as the column in CCC was retaining acid molecules in between runs, potentially distorting the results. It had previously been seen that pigment molecules were staining the inner surface of the CCC column (Sporna-Kucab, 2014). It was important to ensure that the washing procedure after each run was fully removing the acid from the column. The partition coefficient of tolbutamide in HEMWat 17 was measured three times. The first was

made after the CCC machine had been soaked in methanol overnight. The second, after a HEMWat system with formic acid present and third after a HEMWat system with TFA present. In Table 27, it can be seen that the variation in the K_d values for tolbutamide in unadjusted HEMWat 17 is 0.1. This would be considered lower than experimental error and suggests that the porous nature of the Teflon was not leading to acid being retained and altering the partition coefficient.

Table 27 – Following the experimental details in section 2.3.3, the partition coefficient values of tolbutamide were measured in unadjusted HEMWat 17 after a cleaning with a soak in methanol and after using TFA and formic acid in CCC runs.

Compound	Partition coefficient values		
	After methanol soak	After formic acid run	After TFA run
Tolbutamide	0.5	0.6	0.6

3.7.6. Conclusion

In conclusion, the difference in the K_d values obtained by HPLC and by CCC is likely to be due to a change in the pH of the phases during the run. However, due to the difficulties involved with measuring the pH of organic compounds this cannot be confirmed.

There is no significant decrease in the variation of K_d measurement by HPLC achieved by increasing the acid concentration above 0.1%. As this is the standard concentration used in many laboratories, it was decided to maintain this concentration. This low level has the additional advantages of minimising degradation or acid damage. Therefore, it reduces the variability without destroying the molecule.

3.8. Standard Experimental Methodology

The information gained from this chapter was used to define a standard experimental methodology for measuring the K_d values of compounds by HPLC. All further K_d values measured using HPLC were determined using this method.

Six HEMWat systems were made by mass according to Table 1 and replaced into the water bath and left to equilibrate overnight. These were HEMWat 8, 14, 17, 20, 22 and 26. These were chosen as they gave a large polarity range across the whole series. The compound to be studied was weighed into HPLC vials. The solvent systems were removed from the water bath when equilibrated and sampled. The ClogP value of the

compound was used to decide which phase of the HEMWat system was used to dissolve it. A negative ClogP meant that the compounds were dissolved in the lower phase, with compounds dissolved in the upper phase if they have a positive ClogP. Once the phase had been selected, 1.5 ml was saturated with the compound being investigated. To make sure that there were no particulates in the supernatant, the HPLC vials were centrifuged. Then aliquots of 400 µL of compound supernatant were placed into HPLC vials and 1400 µl of the alternative phase was added. The vials were then vortex mixed for 30 seconds and centrifuged at 1400 rpm at room temperature (20°C). An aliquot of 80 µl of the 1400 µl volume phase was placed into 1 ml of ethanol and 320 µl of the 400 µl phase was placed in 1 ml of ethanol. Before the lower phase was sampled the remaining upper phase was removed and discarded. This was done in triplicate for each of the six HEMWat systems. The samples were run on a 10 minute gradient method on the HPLC. The HPLC method details are shown in Table 28. The K_d values of 2-etylanthraquinone should be measured in every experiment to act as an external standard.

Table 28 – HPLC method and conditions

Condition	**Value**	
Column description:	Symmetry C18 4.6 x 75mm 3.5µm Part No. WAT066224 Serial number 016836044316	
Injection volume:	2 µL	
Flow rate:	1 ml/min	
Column temp:	40°C	
Total run time:	10 minutes	
Detector type:	UV DAD	
Detection wavelength used for K_d determination:	275, 295, 325 nm	
Wash vial:	Methanol	
Mobile Phase:		
Time (min)	Percentage 0.1%TFA in water	Percentage acetonitrile
0.00	90	10
2.00	90	10
8.00	20	80
8.50	20	80
8.55	90	10

4. Generating the QSAR models

The computational approach chosen to investigate the prediction of the optimal HEMWat system to achieve separation when using CCC, was Quantitative Structure Activity Relationship (QSAR) modelling. There are many mathematical methods that can be used to train QSAR models. The four being investigated as part of this work, are multiple linear regression (MLR), partial least squares (PLS), support vector machines (SVM) and random forest (RF). Models using each of these methods were built using a diverse training set of compounds to increase the likelihood of the model being able to accurately predict logK_d values. These models were compared to identify the best method for generating the QSAR models.

4.1. Building the training set

To maximise the predictive ability of a QSAR model, the training data set used to build the model must be diverse. The larger the range of end point values used to train the model, the more likely it is that any test compounds will fall into an area of parameter space in which the model has been trained. This increases the probability of an accurate prediction. Assessing this diversity allows the applicability domain of the model to be defined within which the test set must fall to ensure a fair test. A further criterion for the partition coefficient data for the training and test set is that they must be comparable (e.g. measurements taken under the same conditions) and ideally from the same source (Dearden, et al., 2009).

A training set of compounds that had been specifically selected to maximise the range of their octanol/water partition coefficient (logP) values was taken as the initial starting point for the training set (Ignatova, et al., 2011). The compounds had a range of logP values of 5.49 log units. The logP descriptor was used as it is a common lipophilicity parameter widely accepted in the pharmaceutical industry to characterise drugs. However, this did not necessarily mean that this training set provided a good enough coverage of parameter space to build a useful QSAR model. To begin to assess the diversity of the data set, the spread of the five Abraham parameters (see section 1.8.2.1) throughout this potential training set of compounds were examined. As the five parameters: hydrogen bonding acidity (A), hydrogen bonding basicity (B), polarity/polarisability (S), excessive molar refraction (E) and McGowan volume (V) are all numerical, each parameter was divided into bins with the range of 0.2 for easier

graphical representation. The parameter values for each of the potential training set compounds were assigned to a bin and the number of compounds falling within each bin was counted. Figure 22 demonstrates that there were bins that did not contain any compounds and some bins that contain multiple compounds. This lack of spread across the bins may reduce the predictive ability of the model, while the clustering of many compounds into a few bins will lead to a large volume of experimental work that will not enhance the predictive ability of the model.

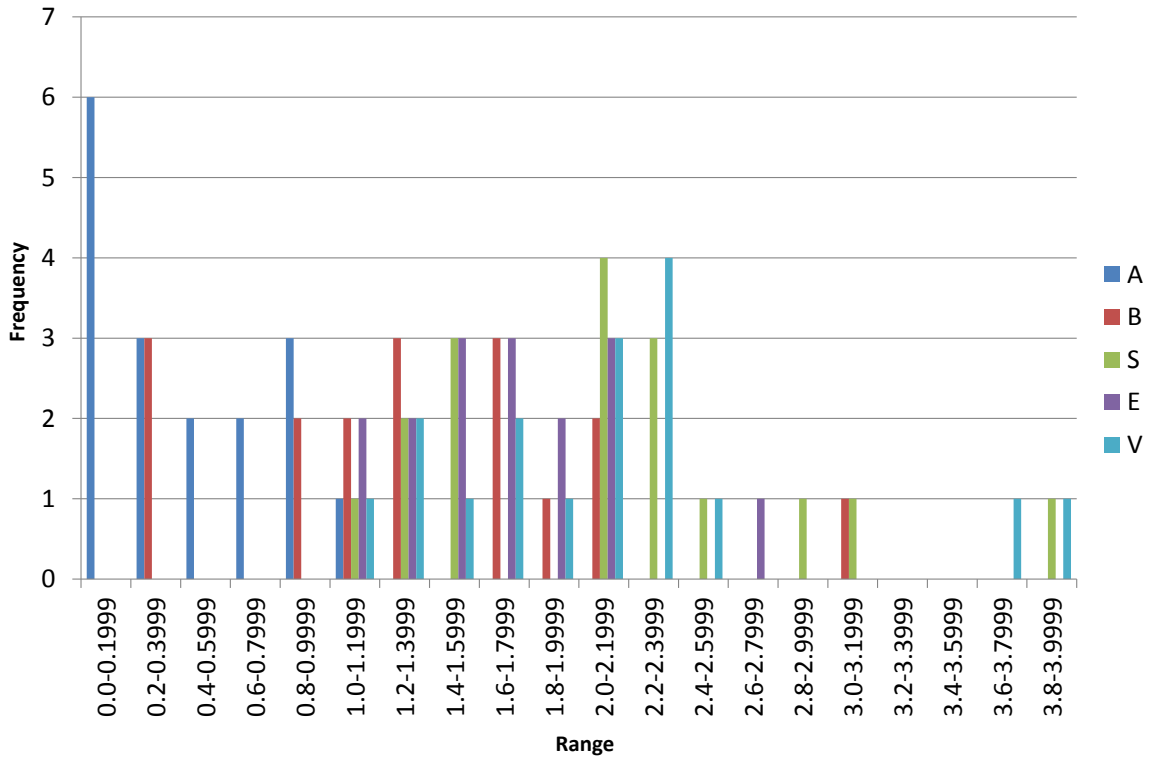


Figure 22 - The coverage of parameter space for the five Abraham parameters of the potential training set as demonstrated by the frequency of compounds within bins each with a parameter ranges of 0.2. The five Abraham parameters are: hydrogen bonding acidity parameter (A), hydrogen bonding basicity parameter (B), polarisability (S), excessive molar refractivity (E) and McGowan volume (V) (Ignatova, et al., 2011).

Having examined Figure 22, it was decided to alter the compound set to try to make it more optimal for QSAR generation. This involved adding in compounds with parameter values that fell into empty bins and removing compounds from bins where there were multiple compounds with similar parameter values. Before a compound was considered for addition to the database, it was assessed against four criteria: availability, price, any hazards posed and its physical state at room temperature. The

latter was a criterion because it was deemed easier to ensure that the HEMWat phase was saturated with a solid and to remove the excess compound. Each of the five parameters were considered individually to ensure there was a compound present in each bin for each parameter to maximise the spread of the training set compounds across parameter space. If more than one compound met the four criteria for one bin, the other parameters for each compound were considered. The compound which fell into the highest number of empty bins was then selected for the training set.

The parameter values for the compounds were obtained from the literature (Abraham, et al., 2009), from a database provided by Professor M. H. Abraham (Abraham, 2013) and from the ABSOLV software (ACDlabs, 2015). These parameter values from Professor Abraham had been experimentally determined, however, with the Abraham parameters from ABSOLV, this was not always the case. It may have been impractical to measure the parameters for certain classes of compounds meaning that they are poorly represented in the ABSOLV database. If it has not been possible to experimentally determine the five parameters, the ABSOLV software will calculate the parameter values. However, it is possible that these calculated parameters contain error.

The parameter space coverage for the final compound set was checked (Figure 23). Despite there being many compounds with a hydrogen bonding acidity (A) value of between 0 and 0.2, there is a broad spread of compounds over a large range, with fewer empty bins. Therefore, based on the Abraham parameters, this training set was considered suitable to be used to build QSAR models for the prediction of K_d values in six HEMWat systems.

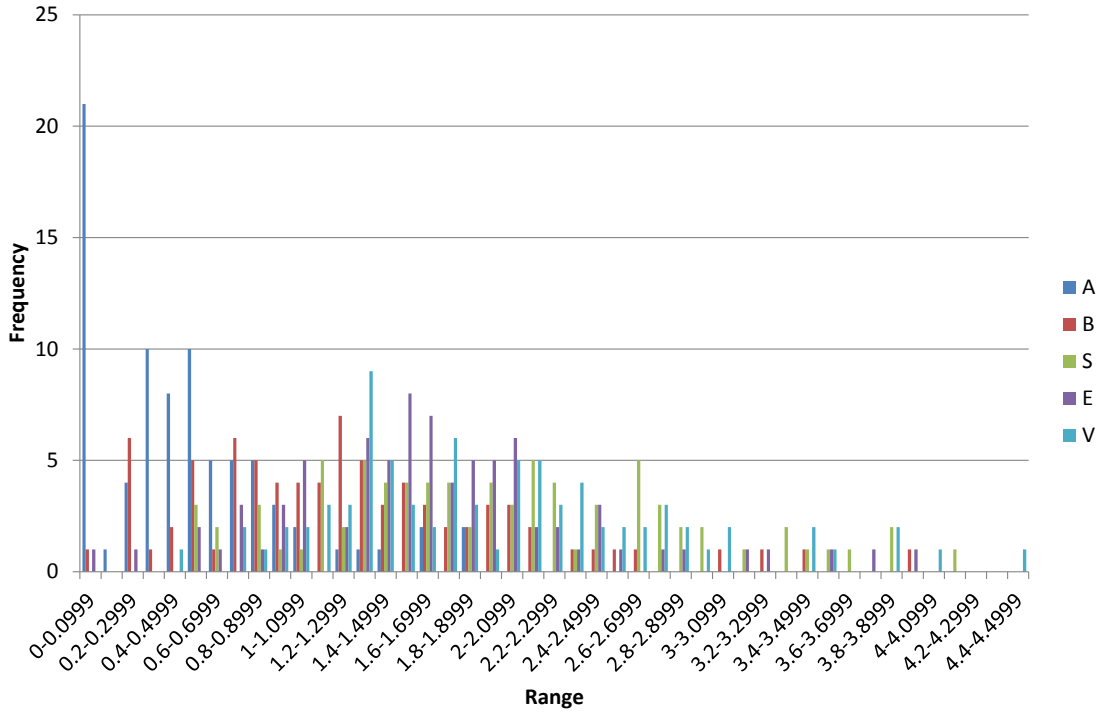


Figure 23 – The coverage of parameter space for the five Abraham parameters of the final compound set as demonstrated by the frequency of compounds within bins each with a parameter ranges of 0.2. The five Abraham parameters are: hydrogen bonding acidity parameter (A), hydrogen bonding basicity parameter (B), polarisability (S), excessive molar refractivity (E) and McGowan volume (V).

4.1.1. Principal Component Analysis

To confirm a good coverage of parameter space had been obtained using Abraham's parameters, principal component analysis (PCA) was carried out on the proposed training set using the commercially available software, SIMCA-P version 13 (Umetrics, Umea, Sweden). The principal components were calculated using all 196 AZ descriptors (listed section 9.3). The graph in Figure 24 demonstrates that a good coverage of parameter space was obtained as there is a good coverage of points across the four quadrants. Therefore, based on the PCA analysis, this training set was used to build QSAR models for the prediction of K_d values in six HEMWat systems.

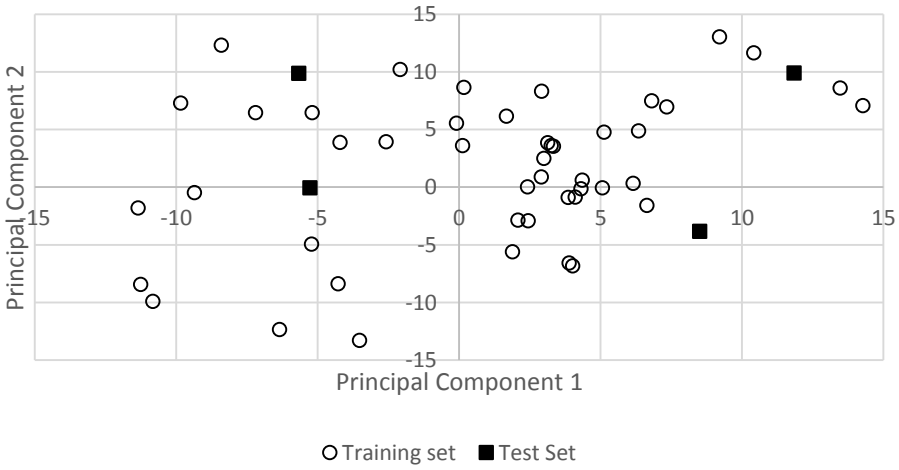


Figure 24 – The principal component analysis (PCA) for the data training set and the four test set compounds of biphenyl, benzoquinone, tolbutamide and quinine.

To test the model, a selection of compounds were withheld from the training set to be used to independently test the QSAR models. There are different ways to select which compounds are used as the test set. A temporal test uses the last 10% of the data that was collected as the test set, whilst a random test set selects 10% of the data at random from anywhere in the data set. A temporal test set mirrors how the model will be used once generated more closely than a random test set. It was decided that the most rigorous way to test the models was to use both types of test set. To allow comparison between mathematical methods and groups of descriptors, a randomly selected test set would first be used to identify the models that are most likely to provide the most accurate prediction of K_d values. Once identified as being likely to provide good predictions, these models would be further validated using a temporal test set. The closeness of the models predicted $\log K_d$ values and the experimental $\log K_d$ values were used to identify the best performing model.

PCA was used to select the compounds for the random test set. As a result, four compounds, which is approximately 10% of the training data set, were selected. These were biphenyl, tolbutamide, quinine and benzoquinone as they represent distinct areas of parameter space to test (represented by black squares on Figure 24). These compounds were well spread across parameter space to ensure the model was tested on a diverse range of compounds.

4.2. QSAR models generated using Partial Least Squares

4.2.1. Method

SIMCA-P version 13 (Umetrics, Umea, Sweden) was used to perform the Partial Least Squares (PLS) regression on the training set of 54 compounds. The initial PLS model was generated using the software tool “Autofit”, which carried out a PLS regression using all 196 descriptors or “Top 14” AstraZeneca descriptors only (listed in section 9.3 and 9.2 respectively). The significance of the descriptors for the original model was assessed using the variance importance for the projection (VIP) values. This is a plot of the VIP values (normalised coefficients) for each of the descriptors allowing the importance of each variable to be assessed. A descriptor with a VIP value greater than 1 indicates that it is significant. Accordingly, any descriptors with a VIP value of less than 1 were removed from the original model and the remaining descriptors were used to generate a second PLS model, again using the “Autofit” tool. Once this model had been built, the same process of removing descriptors with a VIP value of less than 1 was repeated and a third PLS model was built. Please refer to section 2.2.1 for details. The coefficient of determination (R^2), the predictive squared correlation coefficient (Q^2) and the Root Mean Square Error (RMSE) (for definitions see section 1.8.5.1), values were calculated and used to assess the models’ predictive ability. The R^2 value ranges from 0 to 1 with 0 being a very poor fit with the data and 1 being a perfect fit with the data. It is a measure of how well the regression model accounts for variation in the experimental data. It is found by fitting a linear trend line through a plot of the experimentally measured K_d values against the predicted value and adding a linear line of best fit. An R^2 value above 0.78 is considered a good QSAR model (Umetrics, 2015).

The second measure of the predictive ability of the model will be its Q^2 value (see section 1.8.5.4). The higher the Q^2 value, the greater the predictive ability of the model. This is then repeated whilst randomly assigning Y values (in this case, $\log K_d$ values) to compounds in the training set. There should be no relationship as the $\log K_d$ values have no relation to the compounds. If a Q^2 value is above 0.65, it is considered a good QSAR model. This measure of predicted ability was applied to assess models using the training set as there are over 50 compounds.

As the Q^2 value cannot be applied to a test set containing only four compounds, the Root Mean Square Error (RMSE) values was used to assess the results from the external validation. The RMSE value is calculated based on the difference between the predicted and actual values (Equation 57). The lower the RMSE the better the fit of the model. A RMSE value of less than 0.5 indicates that the prediction is acceptable.

Each of these measures was applied to every QSAR model. The best performing QSAR model for the six HEMWat systems was then used to predict the $\log K_d$ values of the four test compounds: biphenyl, benzoquinone, tolbutamide and quinine.

4.2.2. Results

The best model for each of the HEMWat systems was selected on the basis of the R^2 and Q^2 values. A summary of the statistics for these models can be found in Table 29. For five out of the six HEMWat systems, the QSAR models with the best R^2 and Q^2 values were obtained using all 196 descriptors. The exception was HEMWat 8 as the best performing QSAR was obtained using just the "Top 14" AstraZeneca descriptors (listed in section 9.2). The best model for HEMWat systems 8-20 were obtained after all the descriptors with a VIP value of less than one were removed. The best performing models for HEMWat 22 and HEMWat 26 were achieved after the descriptors with a VIP value of less than 1 were removed twice. The performance of the models increase from HEMWat 8 to HEMWat 26 as demonstrated by both the R^2 and Q^2 values.

Table 29 - The details for each of the best performing QSAR models for the six HEMWat systems generated using PLS regression and the descriptor set used to produce it. The R^2 and Q^2 values for the training set of these best performing models are shown along with the number of times the compound with VIP values of less than one were removed.

HEMWat system number	Descriptor Set	No. of times compounds with VIP values of less than 1 were removed	R^2 of the training set	Q^2 of the training set
8	Top 14	Once	0.69	0.66
14	All 196	Once	0.83	0.69
17	All 196	Once	0.81	0.65
20	All 196	Once	0.85	0.61
22	All 196	Twice	0.89	0.80
26	All 196	Twice	0.92	0.86

The best performing models were used to predict the $\log K_d$ values of the random test set of four compounds in each of the six HEMWat systems. These predictions were compared to the experimentally obtained values. The model that made the four predictions which were closest to the experimental values was for HEMWat 22 with an RMSE of 0.27. The predictions from the model for HEMWat 8 were the furthest away from the experimentally obtained values with an RMSE value of 0.67 (Table 30). This is broadly in line with expectation, as the R^2 and Q^2 results for the training set has suggested that the models for the less polar systems (larger HEMWat numbers) were more likely to produce more accurate predictions.

Table 30 – The RMSE values for the test set and the difference between the predicted and experimental values from test set QSAR models generated using PLS by predicting the $logK_d$ values for the diverse test set consisting of biphenyl, benzoquinone, tolbutamide and quinine.

HEMWat number	Average difference between experimental and predicted $logK_d$ value	RMSE test set
8	0.59	0.67
14	0.50	0.60
17	0.34	0.43
20	0.35	0.44
22	0.24	0.27
26	0.42	0.47

The QSAR models from PLS are made up of the sum of coefficients multiplied by the descriptor value, added to the residual constant. The descriptors that the PLS regression identified as significant and the size of the coefficients were examined to gain an insight into the potential mechanisms that are driving the differing partitioning in the six HEMWat systems. To allow a fair comparison of the coefficients for each descriptor identified as significant, the value of the coefficients were normalised. Therefore, the latter represents the change in $logK_d$ (Y) when a specific descriptor (X variable) varies from 0 to 1 while other variables keep at their average. The larger the coefficients on this scale, the higher the correlation between Y and X.

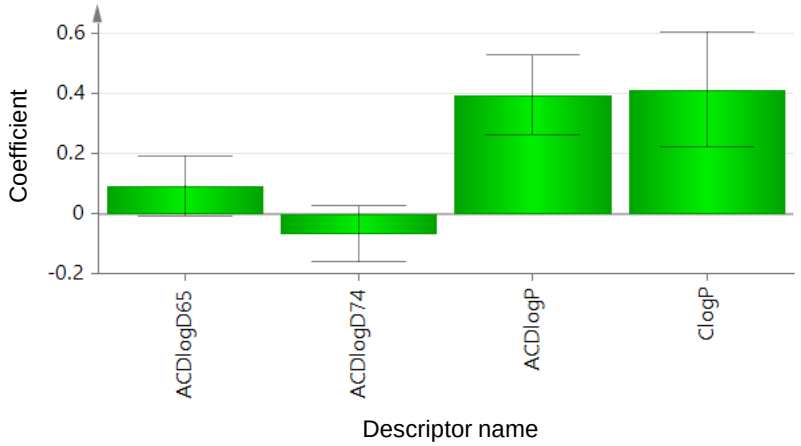


Figure 25 – The descriptors and their normalised coefficients that make up the best performing PLS model for HEMWat 8 which was generated using the "Top 14" AstraZeneca descriptor set. The coefficients are displayed in the form of a histogram to allow direct comparison to ascertain their significance in the relationship between molecular structure and $logK_d$.

All four of the descriptors forming the model for HEMWat 8 are lipophilicity descriptors which describe partitioning in an octanol/water system (Figure 25). As the model is trying to predict partitioning in the HEMWat system which is also comprised of an organic and an aqueous phase, the fact that lipophilicity descriptors are useful in describing this relationship is expected. However, having only four descriptors did not allow enough information to draw conclusions about other factors affecting partitioning.

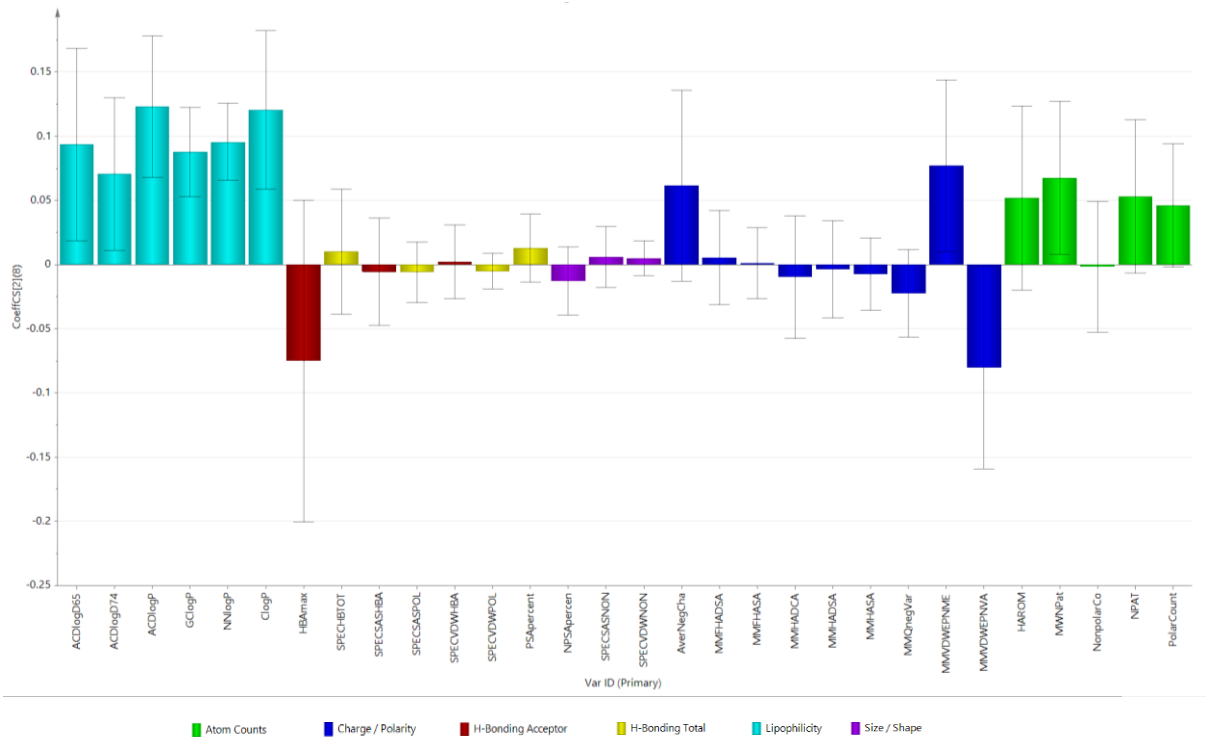


Figure 26 - The descriptors (Var ID) and their corresponding normalised coefficients (CoeffCS) for the best performing model generated using PLS and the 196 descriptor set for HEMWat 8. The full details of the descriptors and their actual coefficient values can be found in the section 9.7.1. The coefficients are displayed in the form of a histogram to allow direct comparison to ascertain their significance in the relationship between molecular structure and logK_d.

Therefore, the descriptors from the second best performing model which was generated using the 196 descriptors were investigated for HEMWat 8 (Figure 26). As expected, the logP descriptors are still significant. However, another of the significant terms is the “HBAmax” term which is a hydrogen bond acceptor term. This may be due to the fact that HEMWat 8 mainly consists of water and ethyl acetate, with the ethyl acetate content of the upper phase being 86.66% and the water content of the lower phase being 96.67% (Ignatova, et al., 2011). Ethyl acetate is a hydrogen bond acceptor whilst water can act as both a hydrogen bond donor and acceptor (Figure

27). Therefore, the hydrogen bond acceptor ability of the solute will have a large effect on the partitioning of the compound. As both solvents have hydrogen bond acceptor ability, compounds with hydrogen bond donor ability are free to partition into either phase. Therefore, descriptors for hydrogen donor ability are not significant.

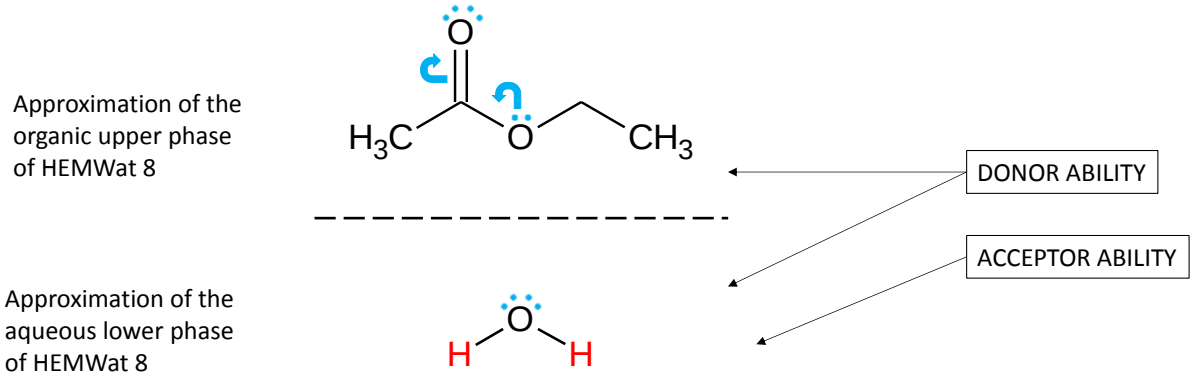


Figure 27 – An approximation of HEMWat 8 with the organic phase containing a majority of ethyl acetate and the aqueous containing a majority of water. The ethyl acetate has a hydrogen bond acceptor ability (the blue lone pairs of electrons). The water has both hydrogen bond acceptor and donor (red hydrogens, H) ability. Any compounds with hydrogen donor ability with preferentially partition into the aqueous lower phase whilst compounds with both hydrogen bond acceptor and donor ability with show no preference.

As the lower phase is mainly water which can act as both a hydrogen bond donor and acceptor, any compounds with hydrogen bond acceptor ability will partition into the lower phase in preference to the upper phase (Figure 27). This is a possible explanation for the presence of the hydrogen bond acceptor descriptor. The significant descriptors from the charge/polarity category are the average negative charge and the mean of the negative charge across the van der Waals surface area. These descriptors are likely to be significant as less negatively charged compounds will be more likely to partition in to the organic phase. The non-polar atom counts, MWNPat and NPat, are significant as compounds containing many non-polar atoms are more likely to partition into the organic phase. The variance of the negative charge across the van der Waals surface area is significant as a large variation across the molecule indicate the molecule is polar which will lead to it partitioning into the aqueous phase.

The coefficients for the best performing model for HEMWat 26 can be found in Figure 29. HEMWat 26 consists mainly of heptane and methanol, with the methanol content of the lower phase being 79.57% and the heptane content of the upper phase being

96.67% (Ignatova, et al., 2011). As heptane has no hydrogen bonds, it cannot act as a hydrogen bond donor or acceptor, meaning that any compounds with hydrogen bonding ability will preferably partition into the lower phase where the methanol has both hydrogen bond accepting and donating ability.

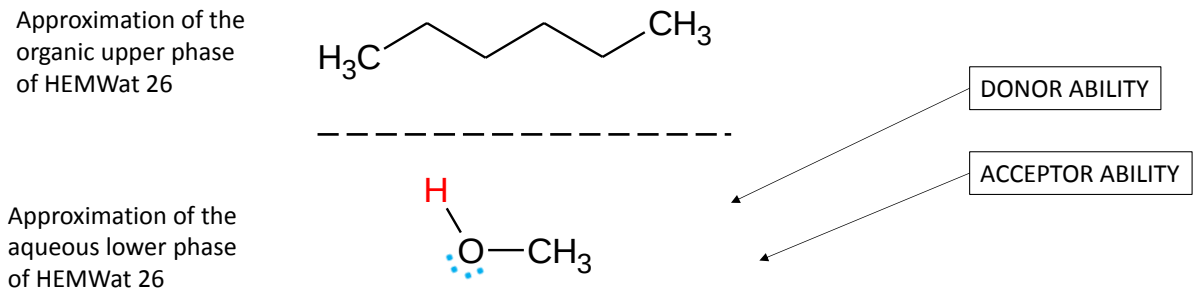


Figure 28 - An approximation of HEMWat 26 with the organic phase containing a majority of heptane and the aqueous containing a majority of methanol. The heptane has no hydrogen bond acceptor or donor ability. The methanol has both hydrogen bond acceptor (the blue lone pairs of electrons) and donor (red hydrogen, H) ability. Any compounds with hydrogen donor or acceptor ability with preferentially partition into the aqueous lower phase.

This offers a potential explanation for the presence of a mixture of hydrogen bond donor and hydrogen acceptor terms in addition to the hydrogen bond total terms in the best performing model. The coefficients are negative, as they were in HEMWat 8, as the compounds that are capable of forming hydrogen bonds will tend to move out of the hydrocarbon layer.

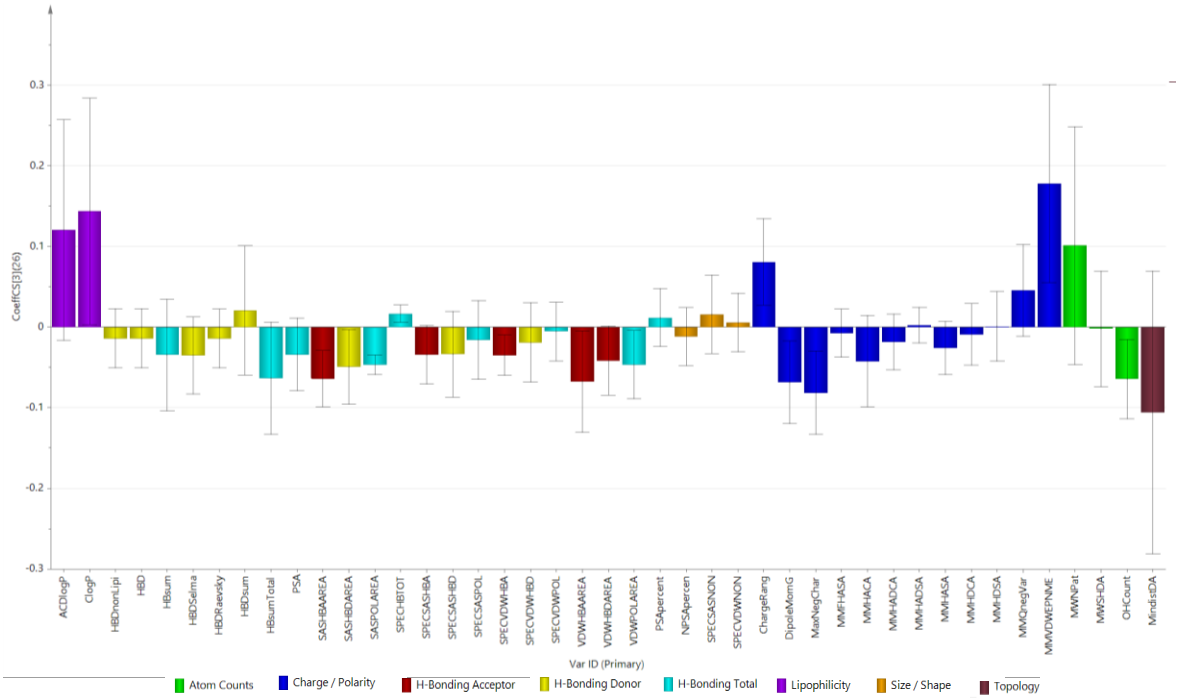


Figure 29 - The descriptors (Var ID) and their corresponding normalised coefficients (CoeffCS) for the best performing model generated using PLS and the 196 descriptor set for HEMWat 26. The full details of the descriptors and their actual coefficient values can be found in the section 9.7.6.

Despite the differences in the hydrogen bond terms for the two models, the most significant descriptors are common to both. The negative charge distribution on the van der Waals surface area, multiple lipophilicity terms, negative charge and non-polar atom counts have large coefficients indicating significance for both of these models suggesting that polarity is a primary driver for the partitioning of a compound in HEMWat systems.

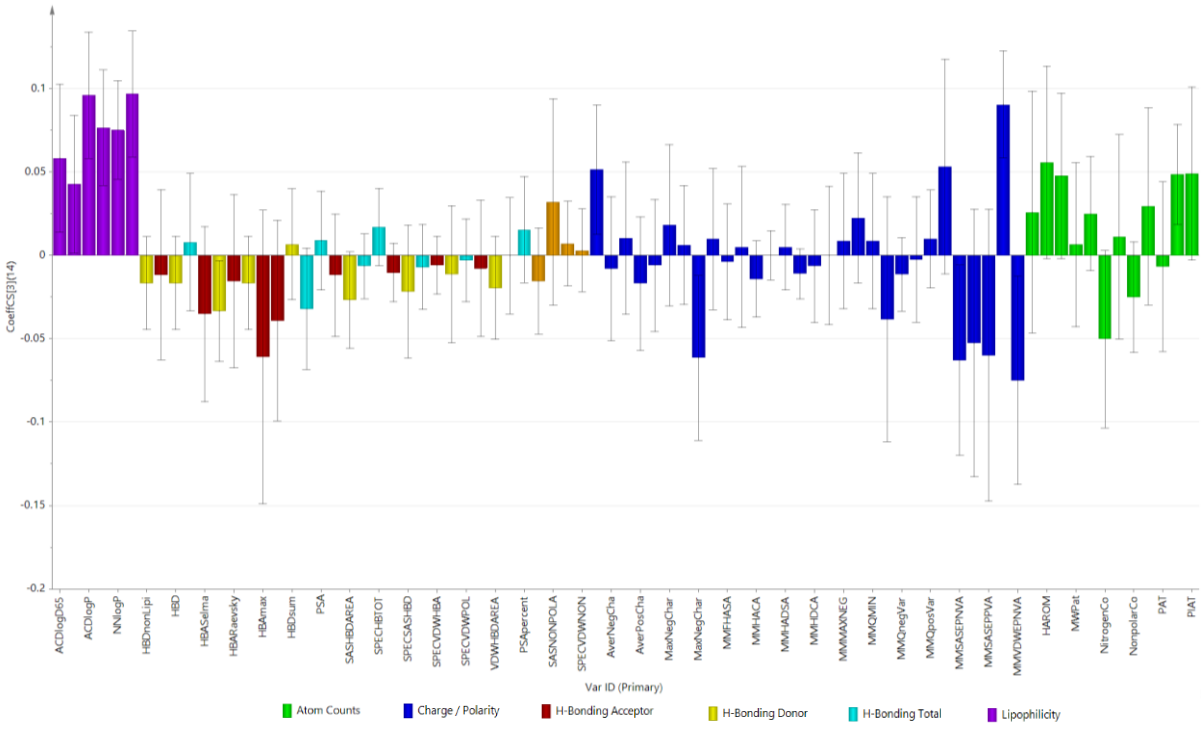


Figure 30 - The descriptors (Var ID) and their corresponding normalised coefficients (CoeffCS) for the best performing model generated using PLS and the 196 descriptor set for HEMWat 14. The full details of the descriptors and their actual coefficient values can be found in the section 9.7.2.

Having examined the models for the most polar and the least polar systems, the four further models were analysed. It was found that, once again, in each of the models the lipophilicity descriptors were significant with five or six logP descriptors in each model (Figure 30, Figure 31, Figure 32 and Figure 33). A further commonality between the models is the mixture of hydrogen bonding acceptor, donor and total terms. This suggests that HEMWat 8 is the only system to have a low enough hydrocarbon content for hydrogen bond donating ability not to play a role in the partitioning of a compound.

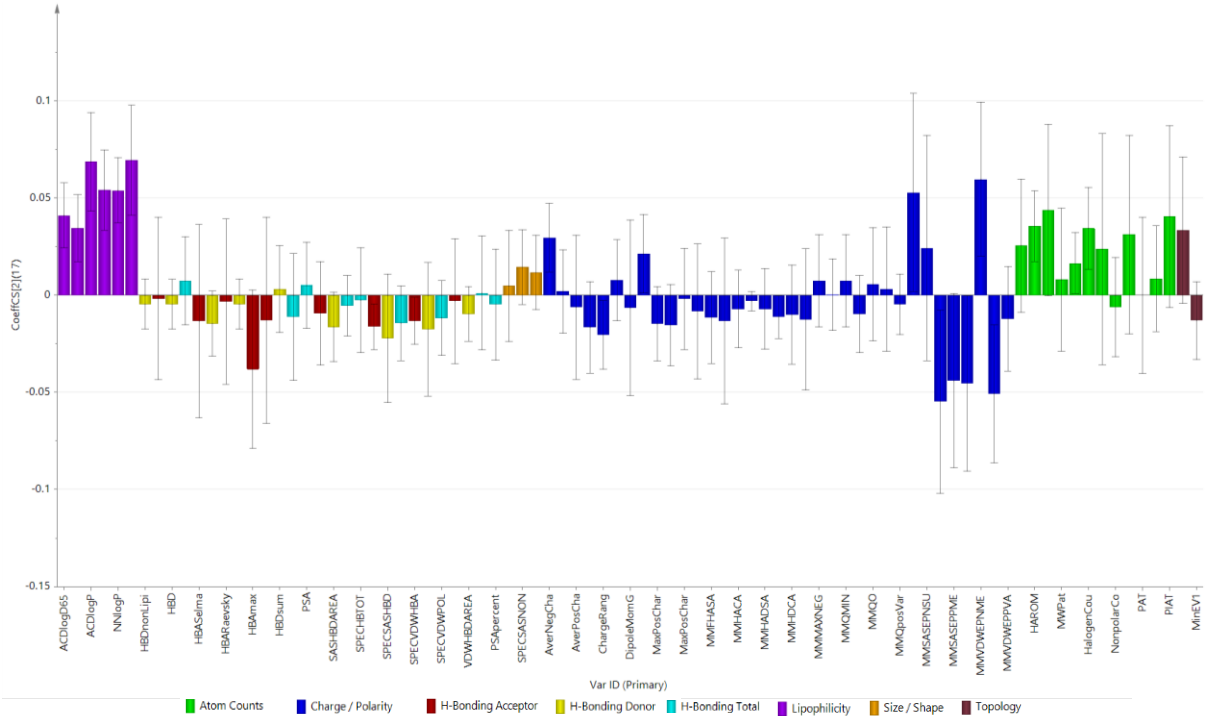


Figure 31 – The descriptors (Var ID) and their corresponding normalised coefficients (CoeffCS) for the best performing model generated using PLS and the 196 descriptor set for HEMWat 17. The full details of the descriptors and their actual coefficient values can be found in the section 9.7.3.

In addition, some of the most significant descriptors across these four models represent the variance of charge across the molecule. This suggests that the polarity of the molecule is a significant factor in all six of the HEMWat systems. The size of the coefficient for these polarity terms increases from HEMWat 14 to HEMWat 22. This suggests that as the hydrocarbon content of the system increases, the polarity of the model becomes most important as the organic phase becomes more non-polar. This change is also reflected in the increase in size of the coefficients for the hydrogen bonding terms.

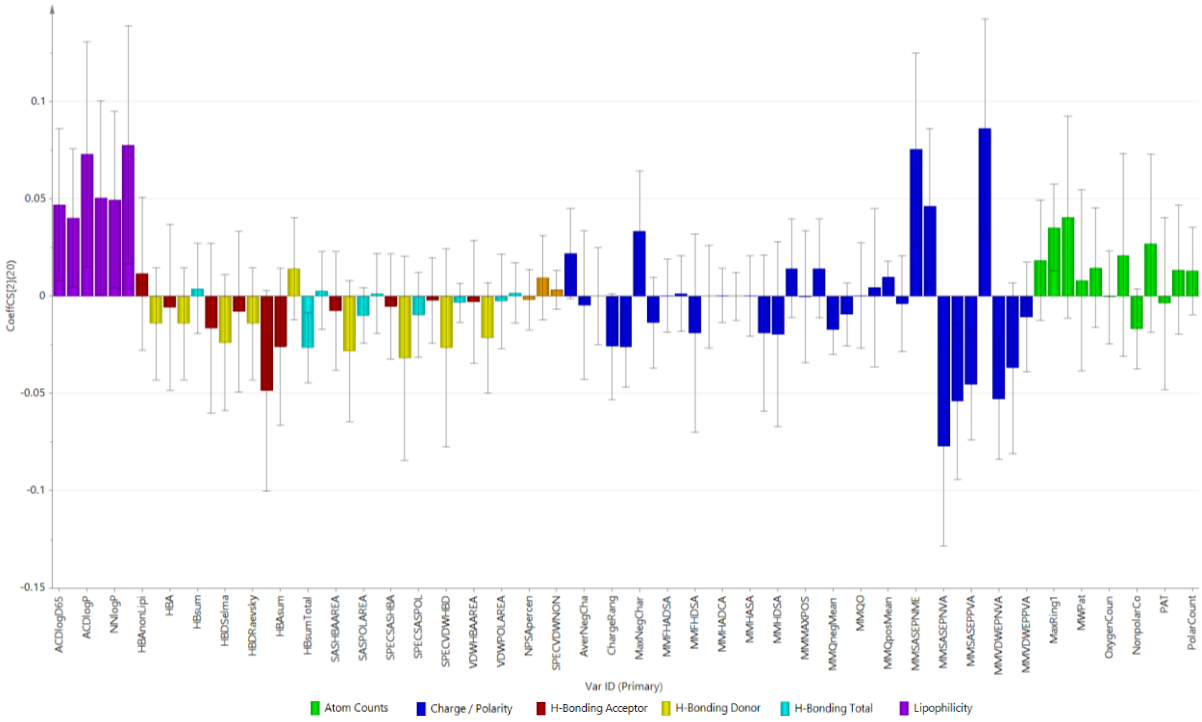


Figure 32 – The descriptors (Var ID) and their corresponding normalised coefficients (CoeffCS) for the best performing model generated using PLS and the 196 descriptor set for HEMWat 20. The full details of the descriptors and their actual coefficient values can be found in the section 0.

Conversely, the coefficients for the descriptors that fall into the category of atom counts, decrease as the HEMWat number of the system decreases. As the majority of these descriptors represent non-polar atom counts, this suggests that as the hydrocarbon content of the organic phase increases the hydrogen bonding ability and polarity terms dominate the non-polar atom counts.

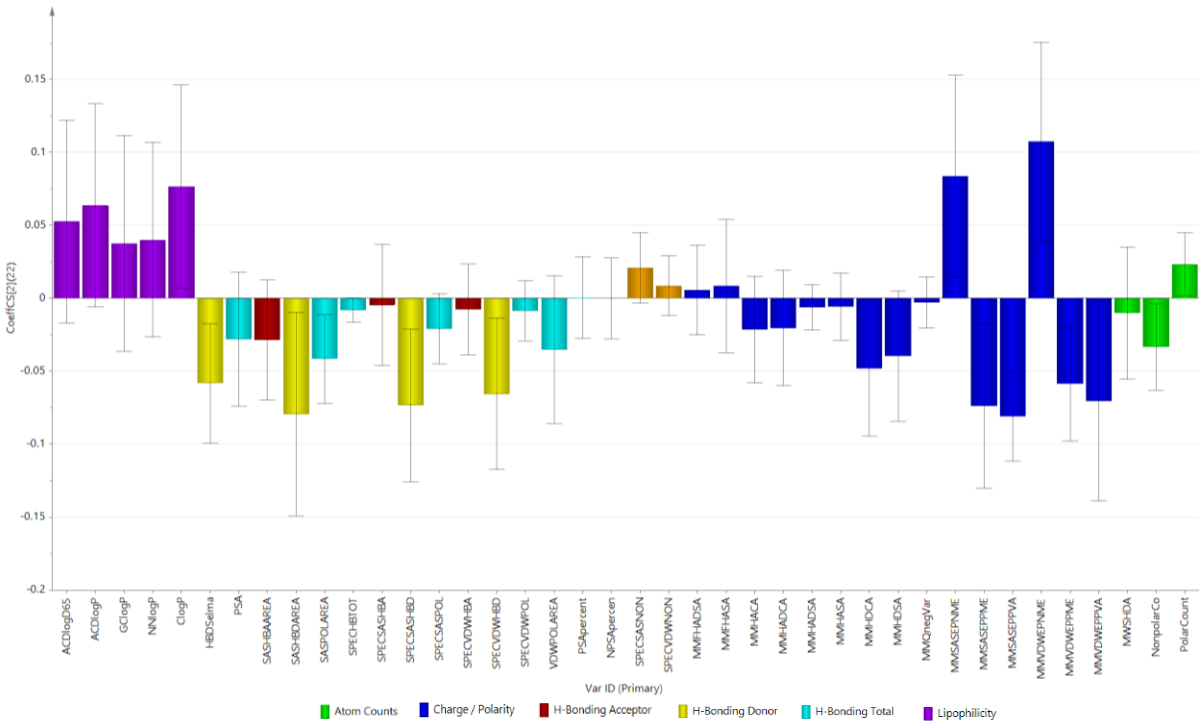


Figure 33 – The descriptors (Var ID) and their corresponding normalised coefficients (CoeffCS) for the best performing model generated using PLS and the 196 descriptor set for HEMWat 22. The full details of the descriptors and their actual coefficient values can be found in the section 9.7.5.

4.3. QSAR models generated using Multiple Linear Regression

The second regression method examined was multiple linear regression (MLR). The 196 molecular descriptors obtained from C-Lab for each molecule provides a large amount of diversity in descriptors, which increases the likelihood of a QSAR model being able to accurately describe the relationship between $\log K_d$ in HEMWat and molecular structure. However, as there are only 54 experimental measurements, using 196 descriptors increases the likelihood of overfitting. A model that is overfitted will appear to model the relationship between the X (in this case, $\log K_d$) and Y (in this case, descriptors) well, but it would have very limited predictive ability. To have the best chance of avoiding overfitting, Dearden et al. recommended that the ratio of the number of training set compounds to the number of descriptors should be a minimum of 5:1 (Dearden, et al., 2009). It was also suggested that no more than six descriptors are used to develop of QSAR model using MLR, to allow understanding of the mechanism of the property being modelled (Dearden, et al., 2009). To guard against this phenomenon, more than one combination of descriptors were used to generate multiple QSAR models so that their predictive ability could be assessed.

4.3.1. Method

The Multiple Linear Regression (MLR) analysis was carried out using the commercially available software JMP 10.0 (SAS Institute, Cary, NC, USA) on the training set of 54 compounds. Several QSAR models were generated using five different combinations of descriptors. A stepwise linear regression was used to select the significant descriptors. This was manually checked by assessing the significance of each descriptor using their *p*-values, with descriptors removed until all the descriptors had a *p*-value of less than 0.05. From these QSAR models, the R^2 and RMSE of the training set were compared. The acceptance criteria for the training set statistics was a R^2 value greater than 0.78 and an RMSE value less than 0.5. However, the ultimate test of a model is its ability to predict test set values.

As it is known that models produced using MLR are prone to overfitting (see section 1.8.3.1.1), the best performing QSAR models were externally validated. By comparing the predicted $\log K_d$ values of the four test compounds with the experimental $\log K_d$ values, this can be assessed. Any model with training set statistics suggesting that accurate predictions will be made, but producing poor predictions when externally validated, is deemed overfitted. Please refer to section 2.2.2 for full details.

4.3.2. Results

The best performing QSAR models from each of the five sets of descriptors (section 2.2.2) were compared for each of the HEMWat systems. Five out of the six best performing QSAR models were obtained using the “Top 14” AstraZeneca descriptors set, with the exception being HEMWat 8. This was as expected, as using all 196 descriptors was likely to increase overfitting.

Table 31 – The best performing MLR models for each of the six HEMWat systems selected on the basis of the R^2 and RMSE values for the training sets. ACDlogP is calculated as the octanol/water partition coefficient for the neutral species, ClogP is a predicted octanol/water partition coefficient from Daylight/Biobyte, SIC is the structural information content of zero order, HBD is the Lipinski number of hydrogen bond donors = number of OH+NH, RingCount is the number of rings (smallest set of smallest rings), VOL is the Gaussian volume and PSA is the polar surface area (Van der Waals radius surface, summed over all N, O and attached hydrogens, 1-3 overlap correction.)

HEMWat system number	Coefficient	Descriptor	Coefficient	Descriptor	Coefficient	Descriptor	Intercept
8	0.750	ACDlogP	6.596	SIC	-	-	-2.088
14	0.614	ACDlogP	-0.303	HBD	-0.314	RingCount	0.211
17	0.597	ACDlogP	-0.367	HBD	-0.319	RingCount	-0.687
20	0.567	ClogP	-0.322	HBD	-0.004	VOL	-1.212
22	0.539	ClogP	-0.395	HBD	-0.004	VOL	-1.130
26	0.281	ClogP	-0.388	HBD	-0.015	PSA	-0.978

In every system, a lipophilicity term has been identified as significant in the MLR equation as had been seen in the PLS models, along with the hydrogen bong donor term (HBD) being present in five of the six systems (Table 31). The exception was the model for HEMWat 8 which was also identified as the exception using PLS. The selection of the descriptor based on the number of rings in the more polar systems, can be explained as compounds with rings would be more likely to partition in to the organic layer as this is related to the lipophilicity of a compound. However, it was not identified by the PLS models as significant. The PSA descriptor which is a hydrogen bonding term was also selected by PLS. Although related, it is unclear why the VOL and the SIC descriptors have been found to be significant.

The best performing models were identified on the basis of their training set statistics (Table 32) with the QSAR models for HEMWat 17 to 26 producing very similar R^2 values with a range of 0.78 to 0.80. The model for HEMWat 14 had a higher R^2 value of 0.87 and the model for HEMWat 8 had a lower R^2 value of 0.71. This is worth noting as the model for HEMWat 8 was produced using all 196 of the descriptors and seems to add weight to the argument that the model has been overfitted. The RMSE value

for HEMWat 14 was the lowest at 0.47 with and the highest for HEMWat 8 at 0.79. This reflected the pattern observed from the R^2 values obtained using the training sets. However, the RMSE for HEMWat 17 and for HEMWat 14 were practically the same (0.48, 0.47), which was unexpected due to the superior R^2 value of HEMWat 14. The RMSE values for HEMWat 20, 22 and 26 are very similar which confirmed the pattern demonstrated by the R^2 values.

Table 32 – The best performing QSAR model for each of the six HEMWat systems generated using MLR and either 196 descriptors or 14 descriptors. The R^2 and RMSE data from the training set was used to select the best performing model. The RMSE statistics for the test set have been used to assess how well the models performed when externally validated.

HEMWat system number	R^2 training set	RMSE training set	RMSE test set
8	0.71	0.79	0.80
14	0.87	0.47	0.55
17	0.80	0.48	0.32
20	0.79	0.56	0.29
22	0.79	0.57	0.33
26	0.78	0.57	0.55

The importance of external validation has been previously noted (see section 1.8.5.3), so the best performing QSAR models for each of the six HEMWat systems were used to predict the $\log K_d$ values of the four test set compounds (Table 32). The models for HEMWat 17, 20 and 22 all produced good predictions, as demonstrated by their RMSE values below 0.5 for the test set. The RMSE data from the test set for HEMWat 8 was very poor at 0.78, which is above the acceptable level of 0.5. It is worth noting that the RMSE values for the test set for HEMWat 14 and HEMWat 26 are also above the acceptance criteria. This suggests that the models for the less polar systems have avoided overfitting whilst the more polar systems are being affected by it. Despite HEMWat 17 having a lower RMSE for the training set, it still predicts well as demonstrated by the low RMSE for the test set.

4.3.3. Additional combinations of descriptors

The overfitting observed for the model of HEMWat 8 was not unexpected as it was the only model of the six to be generated with the initial set containing 196 descriptors. This increases the possibility that significant descriptors were mistakenly removed due to the high level of noise in the model. The suspected overfitting for the model of HEMWat 14 is unlikely to be due to this as the regression was run with the “Top 14” AZ descriptor set only. However, this suggests that there were not enough descriptors available to fully describe the partitioning in HEMWat systems. Therefore, some additional descriptor sets were identified and used to generate QSAR models from MLR to see if better models could be produced. Further QSAR models were generated by MLR using the top 20 coefficients identified using the PLS (see section 4.2.2) and the PLS top 20 coefficients combined with the five Abraham (section 1.8.2.1). The lists of descriptors can be found in section 9.5 and 9.6. The descriptor sets used to produce the best QSAR model for each of the six systems can be found in section 9.8.

Table 33 – The best performing MLR models determined using the R^2 and RSME values for the training set. ACDlogP is calculated as the octanol/water partition coefficient for the neutral species, ClogP is a predicted octanol/water partition coefficient from Daylight/Biobyte, HBD is the Lipinski number of hydrogen bond donors (number of OH+NH), RingCount is the number of rings (smallest set of smallest rings), MWNPat is the Proportion of MW accounted for by the excess of non-polar atoms (by number), MaxNegCharge_GM is the Maximum negative charge, MM_SAS_EP_N_MEAN is the mean of negative electrostatic potentials on solvent accessible surface, A and B are the Abraham parameters for hydrogen bond acidity and hydrogen bond basicity respectively and SAS_HB_A_AREA and SAS_HB_D_AREA are the solvent accessible surface hydrogen bond acceptor and donor area respectively.

HEMWat system number	Coefficient	Descriptor	Coefficient	Descriptor	Coefficient	Descriptor	Coefficient	Descriptor	Intercept
8	0.875	ClogP	-0.012	MWNPat	-	-	-	-	0.599
14	0.614	ACDlogP	-0.303	HBD	-0.314	Ring Count	-	-	0.211
17	0.597	ACDlogP	-0.367	HBD	-0.319	Ring Count	-	-	-0.687
20	0.415	ClogP	3.409	MaxNeg Charge_ GM	0.052	MM_ SAS_EP _N_ MEAN	-0.718	A	0.416
22	0.486	ClogP	-1.427	A	-0.460	B	-	-	-1.390
26	0.293	ClogP	-0.008	SAS_HB _A_ARE A	-0.012	SAS_HB D_AREA	-	-	-1.038

As had been seen with the PLS models and the previous MLR models, lipophilicity was present in each of the models with the ring count descriptor in the models for HEMWat 14 and 17 also being linked to lipophilicity (Table 33). The model for HEMWat 22 and 26 contains both hydrogen bonding acceptor and donor terms which confirmed their presence in the PLS model and was expected due to the high hydrocarbon content of these systems. Along with the fact that a hydrogen bond donor term had been selected in five of the six models, the two negative charge terms were selected as significant for HEMWat 20, are connected to polarity which is known to be important in partitioning. The selection of the MWNPAt term, which describes the proportion of non-polar atoms in a molecule, as significant in HEMWat 8 was unexpected as this system has the lowest hydrocarbon content of the six systems.

The statistics from the best performing models produced using MLR are shown in Table 34. Using the descriptor set of the Top 20 most significant descriptors selected using PLS, produced a large improvement in the training and test set data for the model for HEMWat 26 when compared to the model produced using the top 14 descriptors set (Table 32). Despite the training set data being similar for the models for HEMWat 8, the test set data shown in Table 34 shows an improvement on the model produced using all 196 descriptors (Table 32). This suggests that the effect of overfitting has been reduced. The models for HEMWat 20 and 22 produced using the five Abraham and Top 20 most significant descriptors selected using PLS descriptor set have improved upon the models produce using the top14 descriptor set. This is likely to be due to the ability of the PLS method to identify the most significant descriptors without being effected by overfitting. This in turn allows MLR to calculate the regression with a lower risk of overfitting.

Table 34 – The best performing QSAR model for each of the six HEMWat systems generated using MLR using all of the different combinations of descriptors. The R^2 and RMSE data from the training set was used to select the best performing model. The RMSE statistics for the test set have been used to assess how well the models performed when externally validated.

HEMWat system number	R^2 values for the training set	RMSE values for the training set	RMSE values for the test set
8	0.72	0.78	0.62
14	0.87	0.47	0.55
17	0.80	0.48	0.32
20	0.85	0.46	0.16
22	0.83	0.52	0.37
26	0.86	0.45	0.41

4.4. QSAR models generated using Random Forest

4.4.1. Method

The AstraZeneca internal platform AutoQSAR was used to generate Random Forest (RF) models for each of the six HEMWat systems using all 196 AZ molecular descriptors and the training set of 54 compounds. Please refer to section 2.2.3 for details.

4.4.2. Results

The predictive ability of each model was first assessed using R^2 and RMSE values for the training set. The best performing models were then used to predict the $\log K_d$ values of the four test set compounds. As can be seen in Table 35, the R^2 values for the training set are very unsatisfactory because the R^2 values range from 0.00 for HEMWat 26 to 0.18 for HEMWat 8 whilst they must be above 0.78 to be acceptable. These R^2 values suggest that none of the models will be able to give reasonable prediction of $\log K_d$ values. This conclusion is also supported by the RMSE values which are all more than twice the acceptable level of 0.5.

Table 35 – The R^2 and RMSE training set data statistics for the models produced using Random Forest (RF).

HEMWat system number	R^2 values for the training set	RMSE values for the training set
8	0.18	1.15
14	0.03	1.10
17	0.01	1.17
20	0.09	1.19
22	0.08	1.25
26	0.00	1.33

The models were not externally validated as all six of them failed both the acceptance criteria for the R^2 values and the RMSE values. This may be due to the small size of the data set or the rigid nature of the method within the software not allowing for the optimisation of the number of descriptors or the number of trees and subtrees. In conclusion, the six models produced using RF are extremely poor at predicting the relationship between $\log K_d$ and molecular structure.

4.5. QSAR models generated using Support Vector Machines

4.5.1. Method

The AstraZeneca internal platform AutoQSAR was used to generate Support Vector Machine (SVM) models for each of the six HEMWat systems using all 196 AZ molecular descriptors and the training set of 54 compounds. Please refer to section 2.2.4 for details.

4.5.2. Results

The predictive ability of the models was first assessed using RMSE values for the training set. The best performing models were then used to predict the $\log K_d$ values of the four test set compounds. From the RMSE values for the training set data (Table 36), it can be seen that the model for HEMWat 14 is the only one with an RMSE below the acceptance criteria of 0.5 with a value of 0.46. The remaining five methods do not reach this acceptance criteria so it is likely that they will perform poorly when externally validated.

Table 36 – The RMSE data for the training and test set statistics for the models produced using Support Vector Machine (SVM).

HEMWat system number	RMSE values for the training set	RMSE values for the test set
8	0.53	1.98
14	0.46	1.66
17	0.51	1.49
20	0.55	1.36
22	0.60	1.68
26	0.60	1.14

As can be seen in Table 36, all of the RMSE values for the test set are above one suggesting that the models are very poor. This may be due to the small size of the data set or like the RF models suffer from the rigidity of the software not allowing less significant descriptors to be removed or additional Kernel functions to be fitted.

4.6. Assessing the accuracy of the AutoQSAR platform

The inadequate performance of the models from the machine learning techniques, SVM and RF, was unexpected. As they had been produced through the AutoQSAR platform with no user over sight, it was not clear whether this poor performance was a fundamental problem with using machine learning techniques to model partitioning in HEMWat systems, or a software inaccuracy. In an attempt to establish this, the AutoQSAR platform was used to produce PLS models. The previous PLS models had been generated using the SIMCA software which allowed the removal of non-significant descriptors. A comparison between the predictive ability of the PLS

models from AutoQSAR and those from SIMCA was carried out for each of the six HEMWat systems. This was done by calculating R^2 and RMSE values for the training sets for all of the models. If the PLS models from the AutoQSAR platform and the models from SIMCA were comparable, this would suggest a fundamental problem with using the machine learning techniques of RF and SVM to predict partitioning in HEMWat systems. If the PLS models from AutoQSAR performed significantly worse than the models produced from SIMCA, this would suggest a problem with the AutoQSAR software.

It was found that the PLS models from SIMCA performed considerably better than the PLS models from the AutoQSAR platform with all six of the HEMWat systems as demonstrated by the large difference between the R^2 values for the training sets for all six models for each HEMWat system. The largest difference between R^2 values for the training set using the AutoQSAR and the PLS models for the HEMWat 14 with a R^2 value of 0.00 produce by AutoQSAR and an R^2 value of 0.87 produced by SIMCA. The smallest difference was for the model for HEMWat 8 with R^2 values of 0.09 for the model produced using the AutoQSAR platform and 0.72 for the model produced using SIMCA (Figure 34).

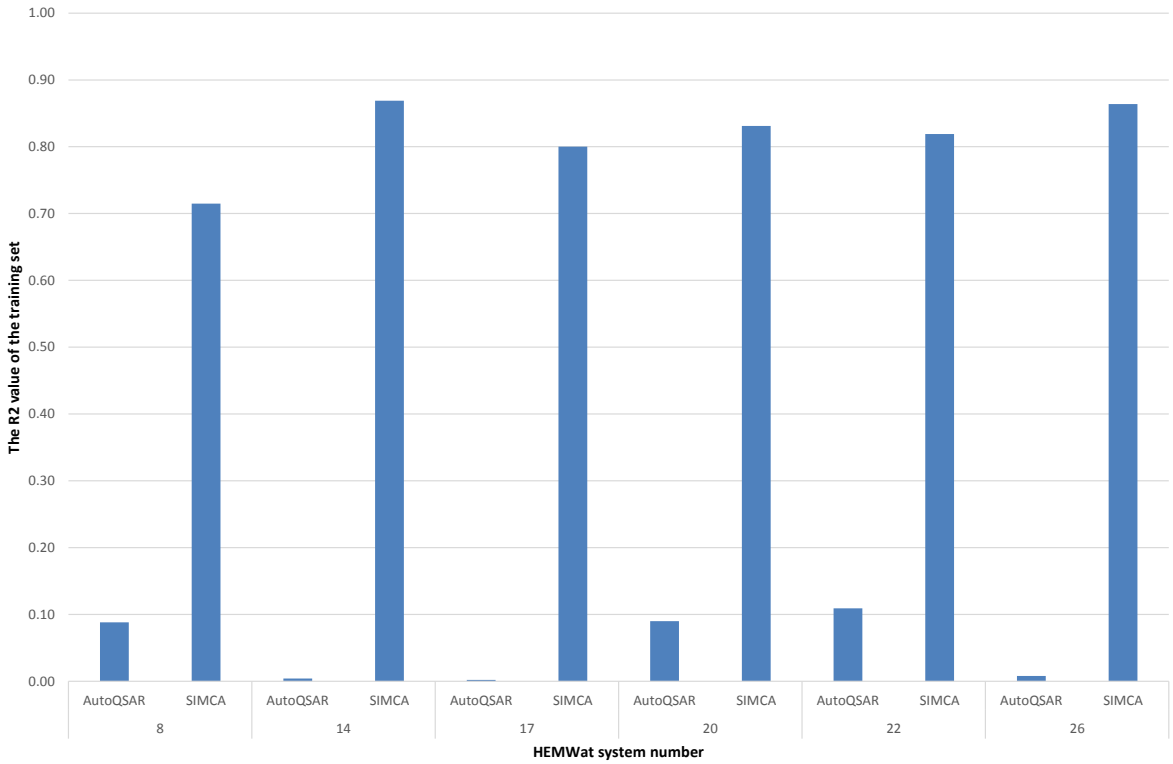


Figure 34 – The R^2 values for the training sets of the PLS models generated by either using the AutoQSAR software or the SIMCA software.

The RMSE data confirms that the SIMCA models outperform the AutoQSAR models (Figure 35). The largest difference between RMSE values for the training set using the AutoQSAR and the PLS models for the HEMWat 26 with a RMSE value of 1.38 produce by AutoQSAR and an RMSE value of 0.45 produced by SIMCA. The smallest difference was for the model for HEMWat 8 with RMSE values of 1.27 generated using the AutoQSAR platform and 0.78 from the SIMCA model.

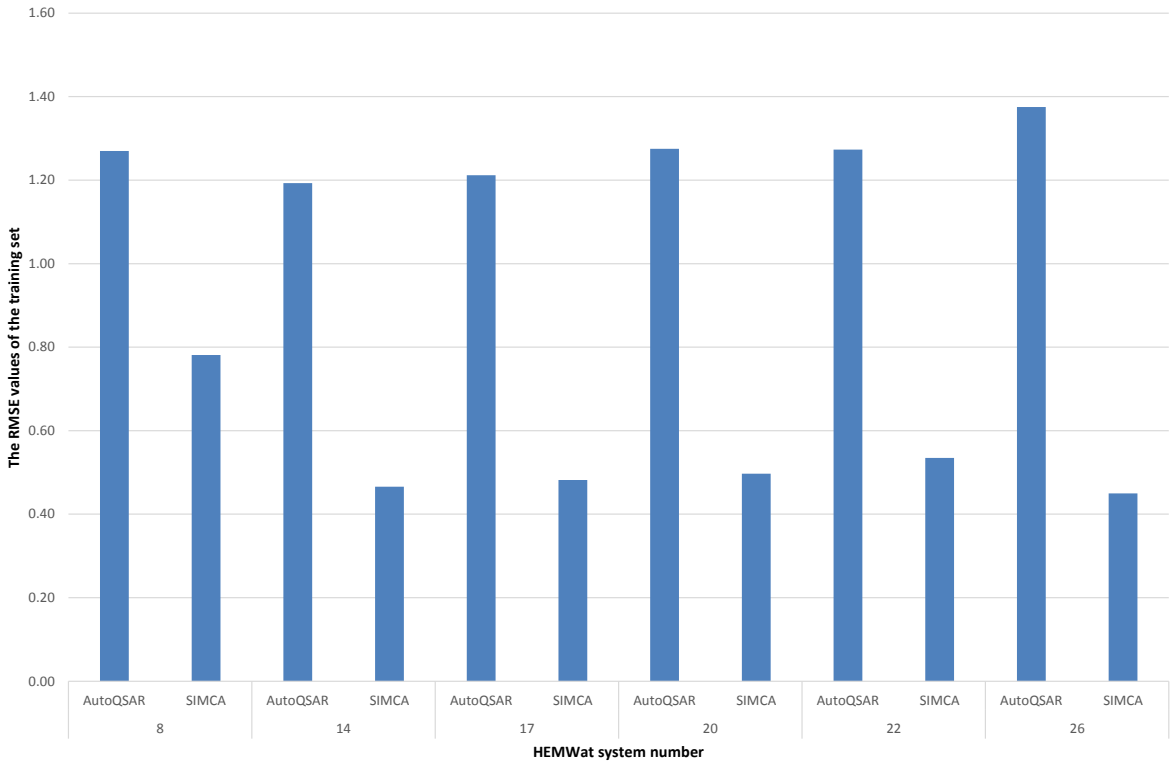


Figure 35 – The RMSE of the training sets of the PLS models generated by either using the AutoQSAR software or the SIMCA software.

This comparison of the PLS models demonstrated that the AutoQSAR platform is not able to produce the optimal model. The fact that the AutoQSAR platform uses all 196 descriptors whereas the SIMCA models only used the most significant descriptors is likely to lead to a poor model. The presence of the descriptors that are not significant in modelling the relationship leads to noise detracting from the accuracy of the model. Both the MLR and PLS models benefited from the ability of the user to manually assess the significance of the descriptors in the model. For MLR this was done through *p*-values and for PLS this was done through the VIP plot. This is likely to be a contributing factor as to why the SVM and RF models are so poor. The small size of the data set could also be causing inaccuracy with some literature recommending that SVM should only be used when there is a large data set available (Ianciuc, 2007). A further problem with the AutoQSAR platform is that the Kernel function is already selected for the SVM model. This is the function that allows the data to be mapped in a higher dimensional space with the aim of increase the likelihood of separating the two classes of data (section 1.8.3.2.2). The radial basis function which is automatically selected by the QSAR software to be the Kernel function may not provide the optimal

fit for the data. Alternative software that allows different Kernel functions to be used could possibly provide a much better fit for the data and therefore, increase the predictive ability of SVM models. This lack of manually optimisation is also likely to have contributed to the poor performance of the RF models. The ability to optimise the number of trees and sub trees may lead to more accurate RF models.

4.7. Comparison of Regression Methods

Given that the models for RF and SVM were so inferior to the models produced using PLS and MLR, they were not tested further. However, the models calculated using PLS and MLR were compared to decide which regression method would be used to build the final models. Once again, both the RMSE and R^2 values were used to perform this comparison. Figure 36 shows the RMSE values for the training and test sets for both regression methods for each of the six HEMWat systems. The QSAR models built using the PLS regression method had lower RMSE values for the training set than the models generated using MLR for four of the six HEMWat systems. The two exceptions were HEMWat 14 and 26 where the lowest RMSE values for the training set were obtained from the models produced using MLR. It is worth noting that there appears to be an overall trend in which the RMSE values decrease as the solvent systems become less polar (higher HEMWat number). However, the RMSE values for the training set statistics from the model produced using MLR are very close to those obtain using the models produced using PLS for HEMWat 17, 20 and 26. Whilst this leads to similar RMSE values for the test set in HEMWat 17 and 26, the QSAR models from MLR for HEMWat 20 outperform the QSAR models from PLS when predicting the $\log K_d$ values of the test set. Another unexpected result was for HEMWat 8. Despite the fact that the RMSE for the training set indicated that the QSAR model from PLS was likely to provide the most accurate predictions, this was actually obtained from the QSAR model from MLR. For HEMWat 14 the QSAR from MLR produced the most accurate predictions as expected. Therefore, the RMSE data shows that the models are very similar.

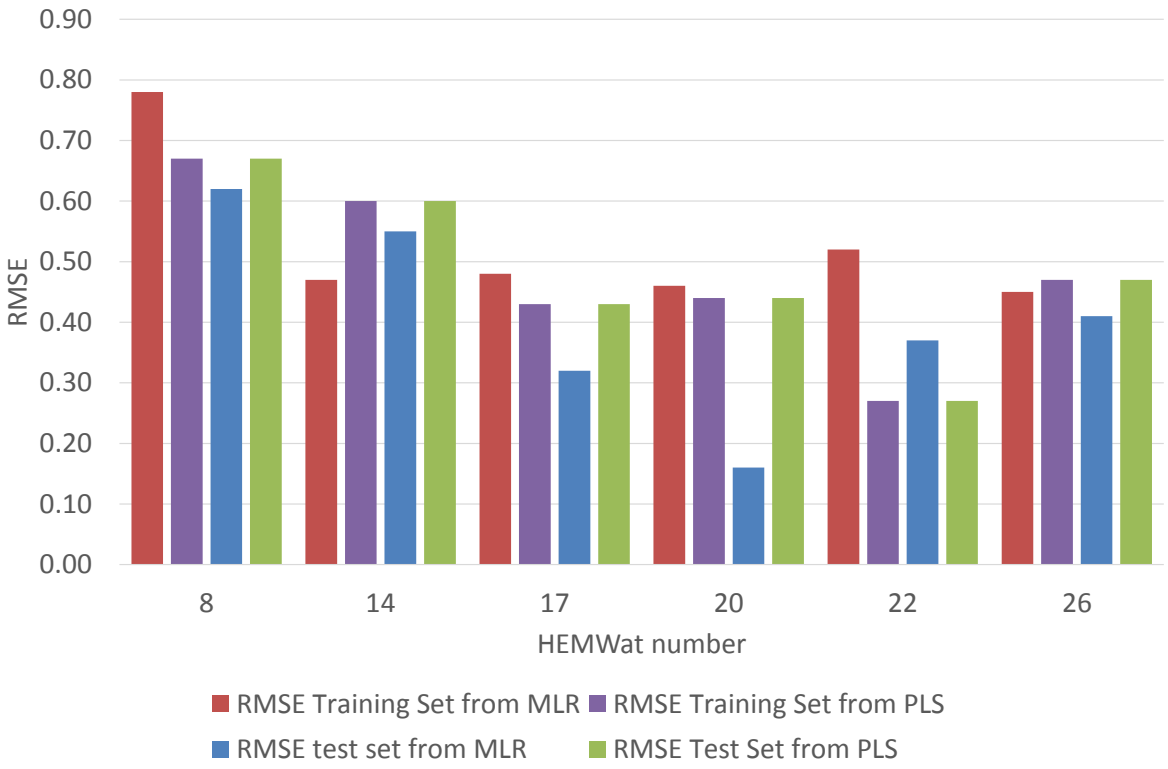


Figure 36 – A comparison of the root mean square error (RMSE) values for the training set and test set for the best performing QSAR models generated by either multiple linear regression (MLR) or partial least squares (PLS).

Figure 37 shows the R^2 values for each of the QSAR models for each of the six HEMWat systems. The R^2 value for the training set suggested that the QSAR model generated using PLS for HEMWat 8 and 14 will produce the best predictions. Whereas, the QSAR generated using MLR for HEMWat 22 and 26 would outperform the QSAR models generated using PLS. This was in agreement with the RMSE data for HEMWat 14 and 26. However, this was contradictory for HEMWat 8 and 22 where the RMSE suggested that MLR would provide the best model. As expected, the R^2 value for the test set shows that the PLS model produced the best predictions for HEMWat 8. At the same time, the QSAR models generated using PLS also predicted the test set $\log K_d$ values most accurately. This was in contrast to the suggestion from the R^2 values for the training set for HEMWat 22 that the best predictions would be obtained using the models built using MLR. The QSAR models generated using MLR produced the best R^2 values for the remaining four systems (HEMWat 14, 17, 20 and 26). This was in keeping with the R^2 values for training set for HEMWat 22 and 26 but had not been anticipated for HEMWat 14. The R^2 values for the training set for

HEMWat 17 and 20 are very close suggesting that there is not much difference in the predictive ability in the models produced by the two regression methods. In contrast, the predictions from the QSAR model from MLR are much better than from QSAR model from PLS for HEMWat 17 and 20.

Figure 37 – A comparison of the R^2 value of the training set and test set for the best performing QSAR models generated by either multiple linear regression (MLR) or partial least squares (PLS).

4.8. Conclusion

Four mathematical methods had been used to produce QSAR models to predict the $\log K_d$ values of compounds partitioned in HEMWat systems. Two were machine learning methods (RF and SVM) and two were multivariate analysis methods (MLR and PLS). The QSAR models produced using the machine learning methods, RF and SVM, had given very poor predictions with every model having a RMSE value for the training set of more than twice the acceptance criteria. Therefore, both these techniques were ruled out. However, it was likely that the poor performance of these models was due to problems with how the AutoQSAR software was applying the methods, not the actual techniques. This would mean that different software employing these two techniques, may be able to produce models that provide good predictions.

The predictions obtained from the multivariate analysis techniques, MLR and PLS, were much more promising, with models for five of the six HEMWat systems producing R^2 values above the acceptance criteria of 0.78. As discussed above, the RMSE and R^2 values for the models generated using MLR and PLS were similar. Despite the similar statistics between the QSAR models, the models produced using MLR have a big disadvantage: MLR is affected by overfitting and the possible effects of collinearity, both of which can lead to inaccuracy. The PLS method is not affected by collinearity and overfitting, however, it does require more descriptors than the models produced using MLR. It was concluded that MLR did not offer any advantages over PLS that outweighed the concerns about possible inaccuracy from collinearity and potential overfitting. Therefore, PLS was chosen as the best option for the QSAR models built to predict CCC/CPC.

5. Experimental validation of the application of the QSAR models

The ultimate test of the utility of the QSAR models is their ability to correctly predict the optimum HEMWat solvent system for the separation of a compound. Two approaches were taken to test the generated QSAR models. The first was to find examples of compound mixtures successfully separated using the HEMWat system from the literature and compare the number of the solvent system employed in the publication to the predicted system obtained from the model. The second was to take a selection of compounds and predict the optimal HEMWat system for CCC separation then run the CCC centrifuge using that system to see if the mixture was split into its constituent parts.

When the structure for each compound is known, each QSAR model can be examined individually to identify the system in which the compounds have a separation factor greater than 1.5, as this is the most likely system to provide separation (see section 1.4). If the structure of every contaminant is not known, no model can fully predict the ideal separation conditions. However, separation is very unlikely to occur if the target compound co-elutes with the non-retained compounds, or elutes late suffering from peak broadening. When there are unidentified contaminants present, separation is most likely to be achieved when the target molecule has a K_d value within the range of 0.5 to 2 (Ren, et al., 2013). As a K_d value of one is in the middle of this range, it was selected as the prediction target for the models.

5.1. Interpolating between the six QSAR models

Having selected PLS as the method of choice for producing the six QSAR models (section 4.8), a strategy was needed for interpolating between the predictions of the six models. This would allow the formation of one overall model that could predict the $\log K_d$ value of a compound in any HEMWat system. The relationship between $\log K_d$ and polarity has been found to be linear and since HEMWat number relates directly to polarity, there is also a linear relationship between $\log K_d$ and HEMWat number (Garrard, 2005). From the data set of 54 compounds, it was found that the six experimentally determined K_d values of 48 compounds had an R^2 values greater than 0.9 whilst the remaining six compounds all had R^2 values above 0.8, when plotted against HEMWat number. By plotting the six predicted values from each of the QSAR

models, a line of best fit was established which was then used to predict the $\log K_d$ values in other HEMWat systems (Equation 58). By solving the equation of the line where x = 0, the HEMWat system in which the compound will have a $\log K_d$ value of 0 (K_d = 1) can be identified and used to predict the optimal system for separation in HEMWat by CCC.

$$\log K_d = gradient \times HEMWat\ number + y\ axis\ intercept$$

Equation 58 – The $\log K_d$ of a compound can be calculated for all 28 HEMWat systems once the gradient and the y-axis intercept from the plot of $\log K_d$ values against HEMWat number from the six models is known.

5.2. Testing the models using literature example CCC separations

From a review of literature and internal reports, compounds with a known structure that had been separated using HEMWat to perform a CCC run, were noted. Only examples that had made use of a HEMWat system from the Arizona table and separated the mixture by exploiting one of the compounds in the mixture having a K_d of one, were used. This permitted the application of the six QSAR models combined using the method in section 5.1. This was to allow a direct comparison of the predictions with the literature examples. The QSAR models were then used to predict the HEMWat system in which the K_d values of the compounds were equal to one. This prediction was then compared to the HEMWat system in which the experimental separation was achieved.

The solvent systems used to measure the QSAR model training set K_d values were made up by mass using the ratios in Table 1. This is in contrast to the majority of solvent systems in the literature which are made up classically by volume or by “mixing on demand”. However, the accuracy of the QSAR model is wholly dependent upon the accuracy of the experimental results. To mitigate against any temperature change altering the volume and therefore the composition of the solvent, the decision was taken to make the solvent systems up by mass. This also had the added advantage that the mass of each solvent was measured on a balance with four decimal places, as opposed to measuring a volume using a pipette or measuring cylinder with no decimal places. However, due to the density differences between the four HEMWat solvents, the percentage compositions of the solvent systems made by mass and those made by volume are different. Table 37 shows the ratios of the solvent systems

by volume when made up by mass. HEMWat systems 8, 14 and 17 are normalised with heptane as 1 and HEMWat systems 20, 22 and 26 and normalised with water fixed at 1. This shows an overall trend of decreasing the overall polarity of the solvent system as the amount of heptane is increased with each system becoming the equivalent of one higher HEMWat number when made by volume. The exception of HEMWat 26 which is only slightly increased i.e. the ratios by volume for HEMWat 8 are approximately 1:6:1:6 which is HEMWat 9.

Table 37 – The volume ratios of the four solvents when HEMWat systems are prepared by mass using the ratios in the solvent system series (Table 1). The ratios for HEMWat 8, 14 and 17 are compared to heptane with this given the value of one. The ratios for HEMWat 20, 22 and 26 are compared to water with this given the value of one. These ratios were compared to HEMWat systems made by volume and the closest system made by volume selected.

HEMWat number when prepared by mass	Heptane	Ethyl Acetate	Methanol	Water and 0.1%TFA	Approximate HEMWat number when prepared by volume
8	1.0	6.8	0.9	6.1	9
14	1.0	1.5	0.9	1.4	15
17	1.0	0.8	0.9	0.7	18
20	2.9	1.1	2.5	1.0	21
22	4.4	1.1	3.8	1.0	23
26	13.2	1.1	11.4	1.0	26

From Table 37, it was concluded that as the model had been trained using systems made up by mass, a prediction from the model that correlated with the literature would be one system number lower than the system used experimentally, as long as the literature system selected was the system in which the target molecule had a K_d value of one.

5.2.1. Testing the models on neutral compounds that had been separated by CCC using a HEMWat system

The first three separations examined consisted of neutral compounds. These mixtures were selected as the majority of the compounds used to train the model had been neutral compounds. The neutral compounds were Astaxanthin, Triptolide, Honokiol, and Magnolol (Figure 38).

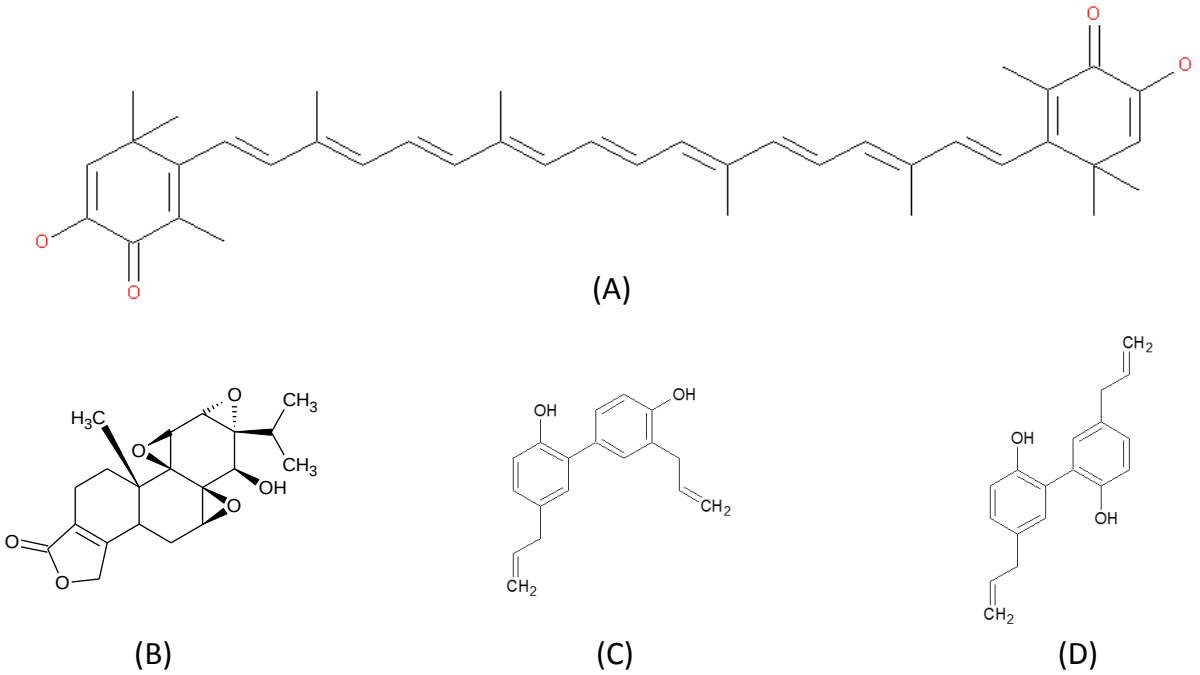


Figure 38 – The molecular structures of Astaxanthin (A), Triptolide (B), Honokiol (C), and Magnolol (D)

When these neutral test compounds were added to the PCA of the training set, all the neutral compounds were within the parameter space (depicted by the ellipse in Figure 39). This indicates that the model has a good chance of being able to provide accurate predictions of the optimal HEMWat system used to perform a separation.

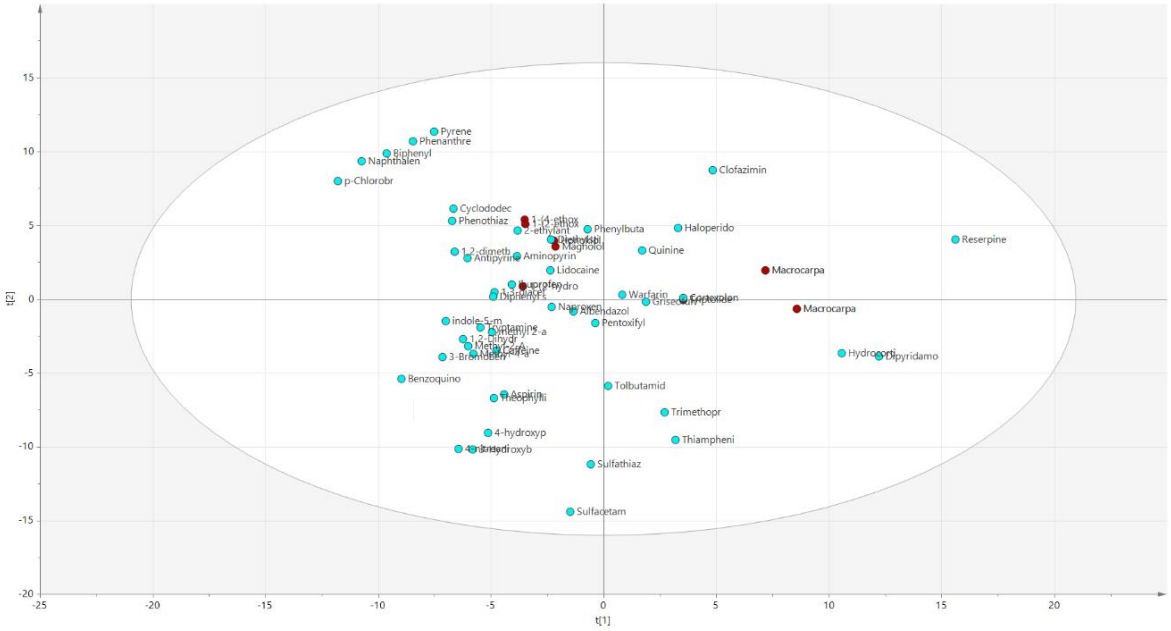


Figure 39 – The principal component analysis (PCA) plot for the training set (light blue circles) with the neutral test compounds from the literature (red circles).

The six QSAR models were used to predict the logK_d values of each of the compounds being tested. The linear relationship between logK_d and HEMWat system number was then used to predict the system in which the compounds would have a logK_d value of zero (a K_d value of one). It is assumed that this solvent system is the most likely to offer separation from contaminant molecules. If the structure of both the compounds being separated is known, then two solvent systems are predicted as possible systems for separation.

As can be seen in Table 38, the QSAR models predicted that the optimal system for separating honokiol and magnolol would be either HEMWat 19 or 20 when they are prepared by mass. As the experimental system used to separate them was HEMWat 21 prepared by volume, the predicted system for magnolol is the HEMWat system used to perform the separation. Therefore, this prediction is very accurate. The system predicted in which honokiol will have a K_d value of 1, is HEMWat 19. It was found that it had a K_d value of 0.3 in HEMWat 21 prepared by volume suggesting the prediction of the compounds having a K_d value of 1 in HEMWat 19 is inaccurate. The next example of a successful separation of neutral compounds was separating out triptolide from a mixture. As the system used to perform the separation was HEMWat 12 prepared by volume, it had been hoped that the model would predict HEMWat 11 prepared by mass. However, the model predicted that HEMWat 14 prepared by mass would provide the optimal conditions for separation. This predicted system is three solvent systems away from the system experimentally used. The next compound investigated was Astaxanthin. HEMWat 25 was experimentally determined as the HEMWat system in which Astaxanthin had a K_d value of one when made by volume. The models predicted the system for optimal separation to be HEMWat 22 when made by mass, which is two system numbers from the experimentally determined optimal system.

Table 38 – The experimentally determined K_d values using CCC and the predicted K_d values for five compounds. The experimentally determined K_d values are from the literature and are for HEMWat systems that were made up by volume. The predictions, however, are for HEMWat systems prepared by mass. To allow a comparison between the model and the experimentally determined value, the prediction must be for the HEMWat system with one less HEMWat number than the system experimentally used.

Compounds	Experimentally HEMWat system (volume)	Experimentally determined K_d value	Predicted HEMWat system (mass)	Predicted K_d values	Difference between predicted and experimentally determined K_d values
Honokiol (Chen, et al., 2007)	21	1.0	20	0.9	0.1
Magnolol (Chen, et al., 2007)	21	0.3	20	0.7	0.4
Triptolide (Hewitson, et al., 2009)	12	1.1	11	3.7	2.6
Astaxanthin (Garrard, 2008)	25	1.0	24	2.4	1.4

A comparison of the K_d values that had been predicted and the experimentally determined K_d values from the CCC was conducted. There is good agreement between the experimental and predicted K_d values for honokiol and magnolol as the difference between the predicted and experimental values are less than 0.5. The predictions for astaxanthin and triptolide are further from the experimental K_d values with a difference greater than one. It is likely that the difficulties in comparing comparisons of K_d values determined by mass and by volume is leading to some inaccuracy. There will also be some inaccuracy added due to the linear fitting of the six QSAR systems.

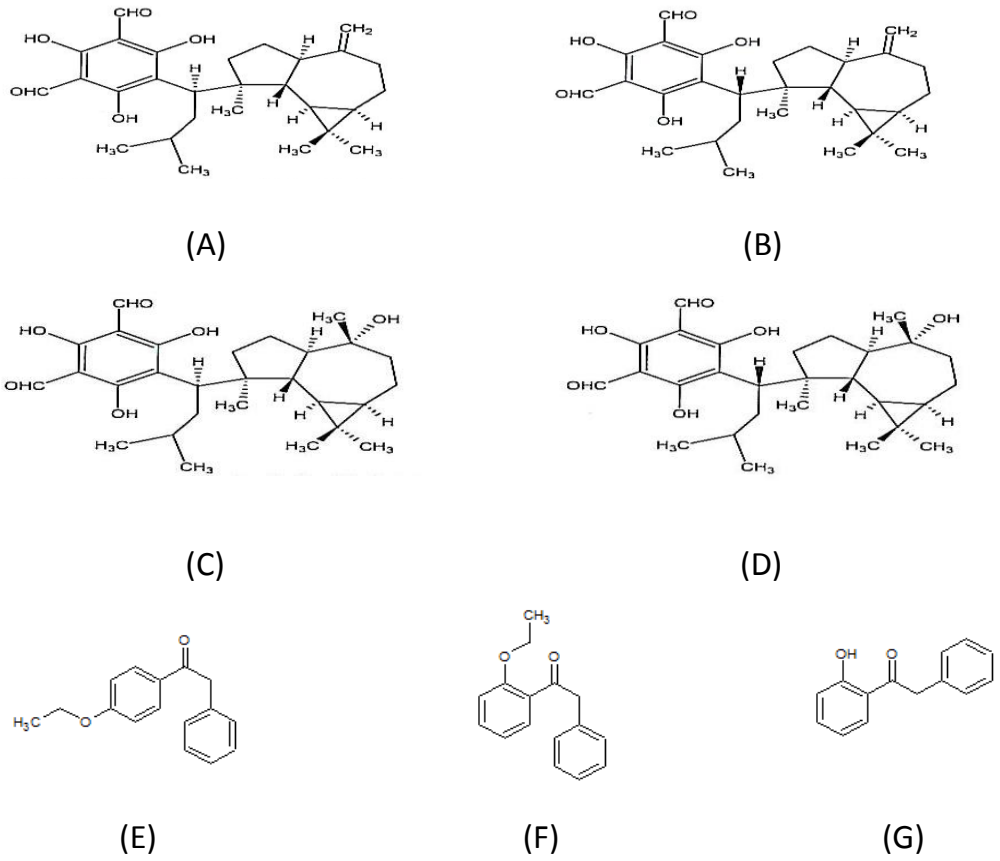


Figure 40 – The molecular structure of macrocarpal C (A), macrocarpal G (B), macrocarpal A (C), macrocarpal B (D), 1-(4-ethoxyphenyl)-2-phenylethanone (E), 1-(2-ethoxyphenyl)-2-phenylethanone (F) and 1-(2-hydroxyphenyl)-2-phenylethanone (G).

The separation of two additional mixtures of neutral compounds were investigated as these mixtures had been separated using gradient methods (Figure 40). As the model is designed to predict an isocratic system to perform separations, it will be unable to predict that a gradient method was used. However, it was found that for both the mixture of macrocarpals and the mixture of phenylethanones, the QSAR models predicted the correct starting point for the gradient (Table 39). However, in both cases the compounds eluted later in the gradient suggesting that this prediction would not allow the separation of these compounds. It is worth nothing that the models do not give any indication that a gradient method would be required or whether the HEMWat system number should be increased or decreased during this gradient.

Table 39 - The HEMWat systems in which the neutral compounds will have a K_d value of one, predicted using PLS, compared to the HEMWat system that was experimentally used to perform the separation.

Compounds	Predicted HEMWat system in which the compounds have a K_d value of one using PLS models	HEMWat system used to perform separation in the literature
Macrocarpal C (Advanced Bioprocessing Centre, 2011)	HEMWat 17	Gradient method HEMWat 17 to 27
Macrocarpal G (Advanced Bioprocessing Centre, 2011)	HEMWat 17	
Macrocarpal A (Advanced Bioprocessing Centre, 2011)	HEMWat 19	
Macrocarpal B (Advanced Bioprocessing Centre, 2011)	HEMWat 19	
1-(4-ethoxyphenyl)-2-phenylethanone (Advanced Bioprocessing Centre, 2012)	HEMWat 20	Gradient method HEMWat 18 to 8
1-(2-ethoxyphenyl)-2-phenylethanone (Advanced Bioprocessing Centre, 2012)	HEMWat 18	
1-(2-hydroxyphenyl)-2-phenylethanone (Advanced Bioprocessing Centre, 2012)	HEMWat 16	

5.2.2. Testing the models on basic compounds that had been separated by CCC using a HEMWat system

As the predictions from the model indicate that the model is capable of predicting the optimal system for separation to within three solvent systems for neutral compounds, the models were then tested against basic compounds in CCC systems including a pH modifier to neutralise any ionisation. Recall bases that formed part of the training set used to build the QSAR models had been neutralised with the addition of a pH modifier of 1% NH_4OH to the water of the HEMWat system.

Using the PLS models to determine the optimal HEMWat system to separate tiamulin showed that the prediction was only one system away from the experimental system which contained 1% NH_4OH as pH modifier/neutralising agent (Table 40). The next example separation attempted was that used to obtain darapladib. Unfortunately, the

predicted HEMWat system was five solvent systems away from the experimental system which is very poor. The reason for this unsatisfactory prediction could be that darapladib is the strongest base in the test set with a predicted pKb of 8.7 (ACDlabs, 2015). As the training set of compounds used to build the models had been trained using neutral molecules and neutralised acid and bases, it is perhaps unsurprising that the models have been unable to produce a more accurate prediction. As can be seen in Figure 41, darapladib is on the edge of the parameter space that had been used to train the model which may also lead to inaccuracy.

The final example separation investigated was spinetoram-J and spinetoram-L. The predicted systems were one solvent system away from the system used experimentally. This was surprising as these two compounds are outside of the parameter space used to train the model (Figure 41) and also the authors did not use a neutralising pH modifier in the CCC experiment. The pKb of spinetoram-J and spinetoram-L is predicted as 9.1 (ACDlabs, 2015) which is similar to darapladib. It is possible that the high accuracy of this prediction is a coincidence.

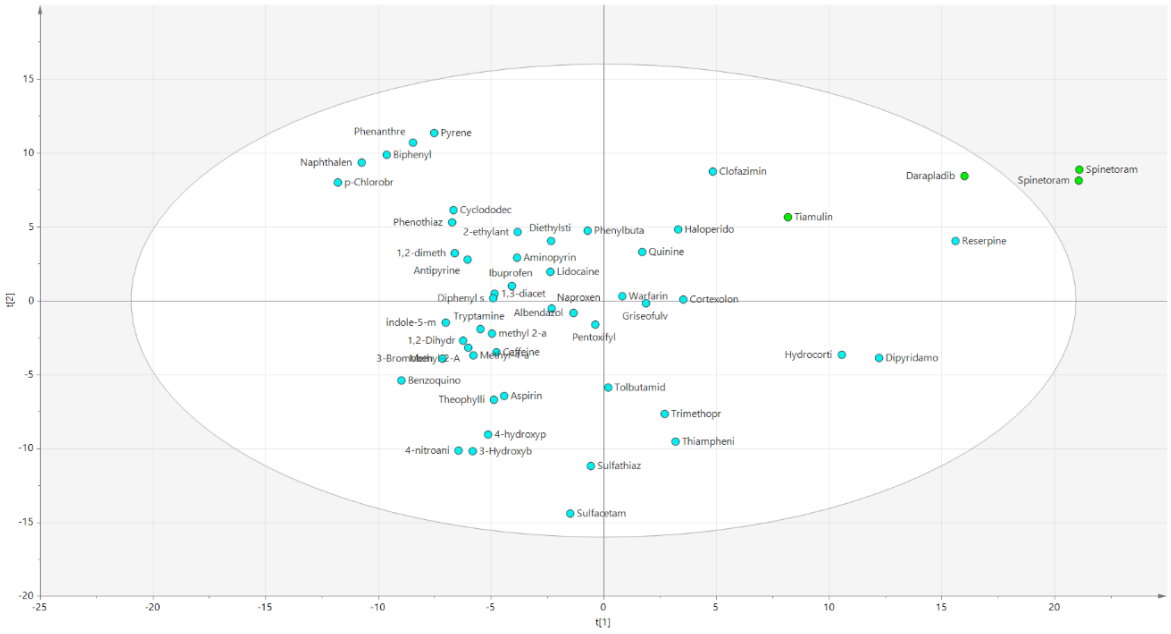


Figure 41 – The principal component analysis (PCA) of the training set (light blue circles) and the base test compounds from the literature (green circles).

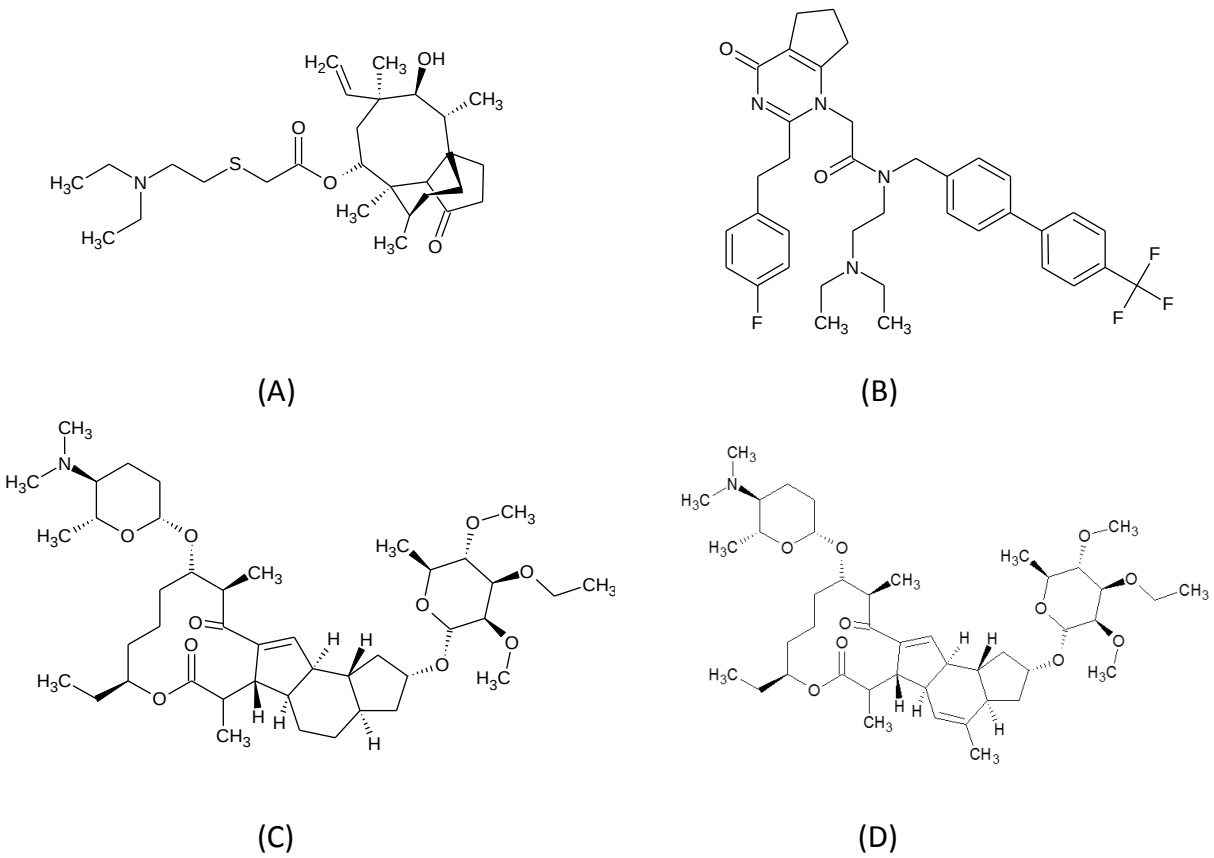


Figure 42 – The molecular structure of tiamulin (A), darapladib (B), spinetoram-J (C) and spinetoram-L (D).

Table 40 - The HEMWat systems in which the compounds will have a K_d value of one predicted using PLS compared to the HEMWat system that was experimentally used.

Compounds	Predicted HEMWat system using PLS models	HEMWat system used to perform separation in the literature
Tiamulin	HEMWat 20	HEMWat 20 + pH modifier of 1% NH_4OH (Advanced Bioprocessing Centre)
Darapladib	HEMWat 23	HEMWat 17.5 (Blackie, et al., 2003)
Spinetoram-J	HEMWat 20 or 22	HEMWat23 (DeAmicis, et al., 2011)
Spinetoram-L	HEMWat 20 or 22	

5.2.3. Conclusion

The predicted solvent system used to separate neutral compounds have been shown to be within three solvent system numbers of the experimental system used. The models were also shown to predict the solvent system used to separate the base,

tiamulin that was run in a system that contain a pH modifier, to within one system. However, the prediction for the more basic darapladib was poor and the prediction for the spinetoram-J and spinetoram-L is likely to be misleading.

Overall, where the test compounds are neutral or neutralised like the training set data, the models are able to predict a solvent system to within three solvent systems of the system used for their successful separation.

5.3. Testing the models ability to predict HEMWat system numbers to successfully separate compound mixtures by CCC

To further investigate the models' predictive ability in identifying the optimum HEMWat system number for CCC compound mixture separation, the six QSAR models were used to predict the logK_d values of each compound in each of the six HEMWat systems for a number of test compound mixtures. As the structure of all of the compounds was known, the system in which the compounds had a separation factor (α) larger than 1.5 was selected, as this suggests that base line separation will be achieved during a CCC run (section 1.4). The test mixtures were selected from those commonly used to test HPLC columns (HICHROM Chromatography Columns and Supplies). The compounds selected had to be reasonably priced, not pose a large hazard and be solid at room temperature.

The identification of the compounds in the CCC chromatogram was conducted by matching the HPLC K_d value with the CCC K_d value, as well as by comparing the spectra of the peaks if required.

5.3.1.1. Separating uracil, phenol, o-terphenyl and triphenylene

The first mixture separated included uracil, phenol, o-terphenyl and triphenylene (Figure 43). The pKa values of uracil and phenol are 9.45 (Drugbank, 2015) and 9.99 (Drugbank, 2015) respectively. This mixture was chosen due to the diverse nature of the compounds.

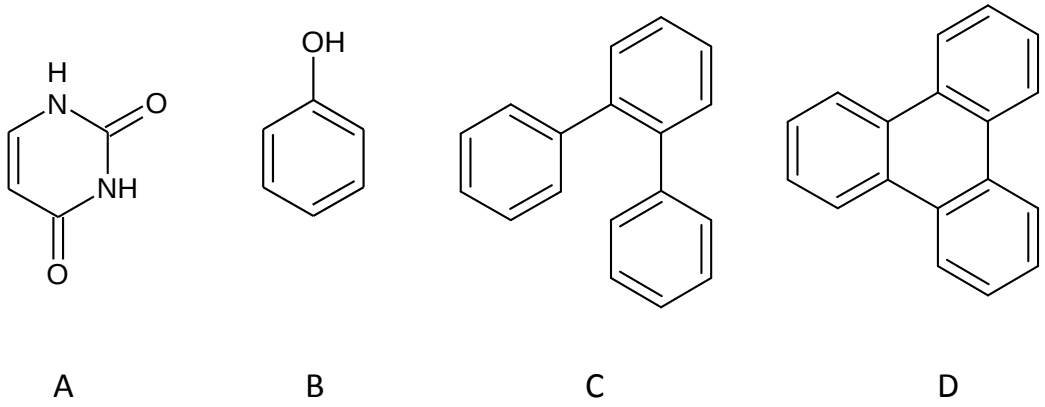


Figure 43 – The molecular structures of uracil (A), phenol (B), o-terphenyl (C) and triphenylene (D).

The predictions shown in Table 41 suggested that it would not be possible to separate all the components of a mixture with a reasonable run time. This is due to the very high K_d values of o-terphenyl and triphenylene in the systems with low HEMWat numbers increasing the likelihood of peak broadening and co-elution. In addition, uracil and phenol have the same K_d value of zero in the systems with high HEMWat numbers, so will co-elute. Therefore, two different systems would be used to separate the mixture. The very non-polar o-terphenyl and triphenylene could be separated using HEMWat 26 ($\alpha = 1.7$), with the predicted K_d values also suggesting that in this system, the uracil and phenol would coelute at the solvent front. The model for HEMWat 26 predicted that the o-terphenyl and triphenylene would form two peaks, the first after one and a half column volumes indicating o-terphenyl and the second after two and a half column volumes indicating triphenylene.

Table 41 – The predicted K_d values for uracil, phenol, o-terphenyl, and triphenylene from the six QSAR models generated using PLS for each of the six HEMWat systems.

Compounds	Predicted K_d values from the six QSAR model for six HEMWat systems					
	8	14	17	20	22	26
Uracil	0.4	0.1	0.0	0.0	0.0	0.0
Phenol	15.8	1.00	0.1	0.0	0.0	0.0
O-Terphenyl	3224.5	629.6	80.1	41.3	2.5	1.4
Triphenylene	6166.0	669.71	107.2	147.9	3.6	2.5

The compounds were dissolve in the upper phase of the HEMWat 26 system which was then used to carry out the CCC run in reversed phase. The experimental details can be found in section 2.3.2. The displaced volume of the equilibrated system was 13 ml, therefore stationary phase retention was 35%.

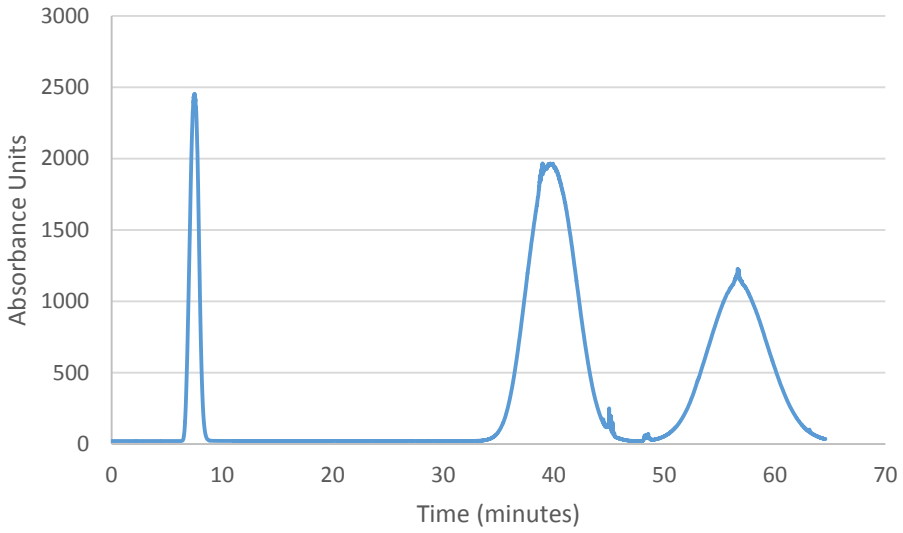


Figure 44 – The CCC chromatogram for the separation of uracil, phenol, o-terphenyl and triphenylene in HEMWat 26 with 0.1% TFA in water replacing the water. The first peak contains uracil and phenol co-eluting at the solvent front (7.6 minutes), whilst the remaining two peaks were o-terphenyl and triphenylene.

As can be seen in Figure 44, the chromatogram from the separation in HEMWat 26 was as predicted with uracil and phenol eluting with the solvent front and o-terphenyl and triphenylene forming two peaks. Measurement of the K_d values for the four compounds in the mixture by either the retention time CCC method or shake flask HPLC showed comparable values. However, comparison of these measured K_d values with those predicted by the model showed that while three of the compounds had predicted well, o-terphenyl had been under predicted. This resulted in the elution order of o-terphenyl and triphenylene being reversed. (Table 42). This was confirmed by the spectra of the CCC peaks. The second peak in the spectra was found to have two λ_{max} values matching the spectrum for triphenylene. The third peak was found to have one λ_{max} value corresponding to o-terphenyl. When the CCC K_d values were compared to the predicted values from the QSAR models, the difference for the values for triphenylene was 0.2. However, the prediction for the o-terphenyl was 2.6 units away. The poor prediction of the K_d value for o-terphenyl may be due to all of the

neutral compounds comprising exclusively of benzene rings fused like triphenylene, whereas the benzene rings in o-terphenyl are joined by single C-C bonds.

Table 42 – A comparison of experimentally determined K_d values from the CCC chromatogram and the predicted K_d values of uracil, phenol, o-terphenyl and triphenylene in HEMWat 26.

Compounds	Predicted K_d values from the PLS QSAR model	Experimental K_d value obtained using retention time on CCC chromatogram	Experimental K_d value obtained using shake flask HPLC method
Uracil	0.0	0.0	0.0
Phenol	0.0	0.0	0.1
O-Terphenyl	1.5	4.1	4.7
Triphenylene	2.5	2.7	2.8

The results above demonstrated that HEMWat 26 could be used to separate o-terphenyl and triphenylene. However, the uracil and phenol had coeluted with the solvent front. The predicted K_d values from the QSAR models suggested that by running the mix in HEMWat 14, these two compounds could be separated whilst maintaining a reasonable run time (Table 41). The predictions for HEMWat 14 suggested that uracil would elute with the solvent front, whilst phenol will produce a peak after one column volume with a separation factor of 10. Any remaining o-terphenyl and triphenylene will be retained on the column.

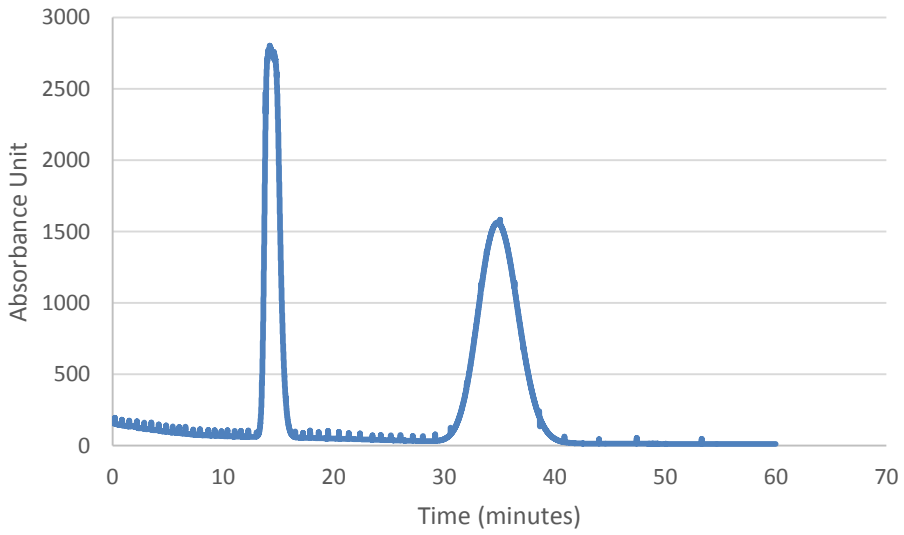


Figure 45 – The CCC chromatogram for uracil, phenol, o-terphenyl and triphenylene in HEMWat 14. The uracil eluted with the solvent front (14.7 minutes) followed by the phenol (35.3 minutes). The o-terphenyl and triphenylene were retained on the column so did not result in a peak.

The compounds were dissolved in the upper phase of the HEMWat 14 system which was then used to carry out the CCC run in reversed phase. The experimental details can be found in section 2.3.2. The displaced volume of the equilibrated system was 12 ml, therefore the stationary phase retention was 40%. The chromatogram in Figure 45 shows the two predicted peaks with uracil eluting with the solvent front and phenol eluting at 35 minutes. Table 43 shows a comparison between the experimentally measured K_d value obtained from the CCC chromatogram and the predicted values from the QSAR model. Although there are differences between the K_d values, the model has still predicted a HEMWat system that separated uracil and phenol.

Table 43 – A comparison of experimentally determined K_d values from the CCC chromatogram and the predicted K_d values for uracil, phenol, o-terphenyl and triphenylene in HEMWat 14.

Compounds	Predicted K_d values from the PLS QSAR model	Experimental K_d value obtained using retention time on CCC chromatogram	Experimental K_d value obtained using shake flask HPLC method
Uracil	0.1	0.5	0.0
Phenol	1.0	2.9	5.7
O-Terphenyl	629.6	Large	-
Triphenylene	669.7	Large	-

5.3.1.2. Separating sulfanilamide, sulfamethoxypyridazine, sulfamethoxazole and sulfapyridine

The next test mixture contained sulfanilamide, sulfamethoxypyridazine, sulfamethoxazole and sulfapyridine (Figure 46) with their pKa values being 10.6 (Drugbank, 2015), 6.7 (Druglead, 2015), 5.6 (Tonnesen, 2004) and 8.43 (Drugbank, 2015) respectively. This mixture was chosen due to the similarity of the compounds.

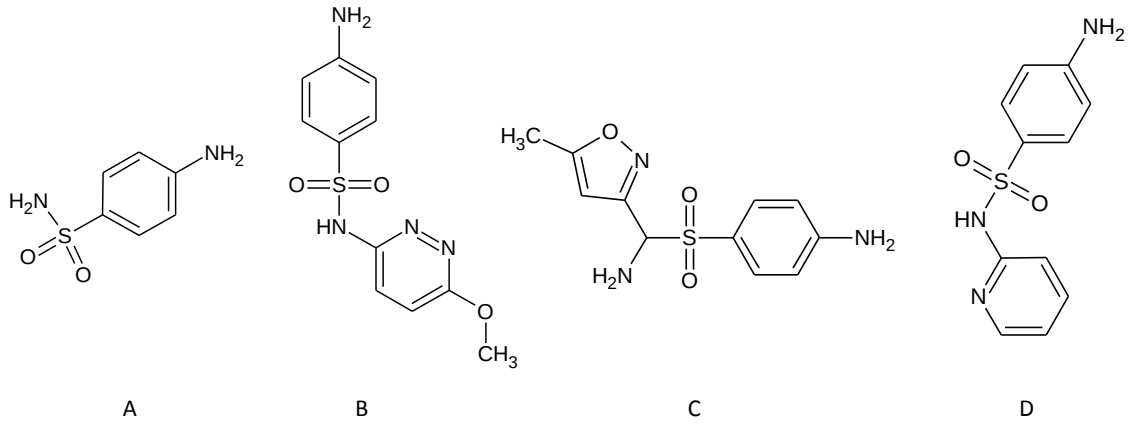


Figure 46 – The molecular structures of sulfanilamide (A), sulfamethoxypyridazine (B), sulfamethoxazole (C) and sulfapyridine (D).

The same method as above (section 5.3.1.1) was employed to select a HEMWat system for the separation of this test mixture. HEMWat 14 was selected as the system that was the most likely to provide the separation of the components of the mixture from the predicted K_d values (Table 44) as the separation factors (α) are above 1.5 in two of the three cases (for sulfapyridine/sulfamethoxazole α = 2, for sulfamethoxazole/ sulfamethoxypyridazine α = 1.5 and for sulfamethoxypyridazine/sulfanilamide

α = 1.4) The HEMWat systems 17, 20, 22 and 26 could not be used as all four compounds would co-elute whilst only one of the three pairs of compounds had a predicted separation factor above 1.5 if run in HEMWat 8. Carrying out the separation in HEMWat 8 also had the added disadvantage of predicted K_d values above two increasingly the likelihood of peak broadening leading to co-elution.

Table 44 - The predicted K_d values for sulfanilamide, sulfamethoxypyridazine, sulfamethoxazole and sulfapyridine from the six QSAR models generated using PLS for each of the six HEMWat systems.

Compounds	Predicted K_d values from the six QSAR model for six HEMWat systems					
	8	14	17	20	22	26
Sulfanilamide	0.8	0.3	0.0	0.0	0.0	0.0
Sulfamethoxypyridazine	3.6	0.8	0.1	0.0	0.0	0.0
Sulfamethoxazole	4.6	1.2	0.1	0.0	0.0	0.0
Sulfapyridine	4.7	1.7	0.1	0.0	0.0	0.0

Interestingly, only two peaks were seen in the chromatogram (Figure 47).

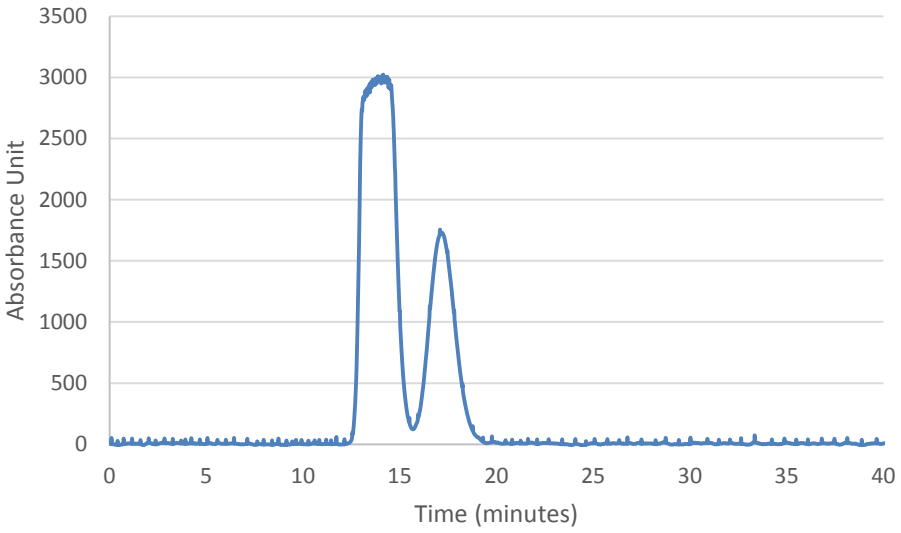


Figure 47 - The separation of sulfanilamide, sulfamethoxypyridazine, sulfamethoxazole and sulfapyridine was separated using HEMWat 14 prepared by mass and left to equilibrate overnight.

As it had not been possible to obtain K_d values using the shake flask HPLC method for all four compounds (Table 45) and the similarity of the spectra of the sulfanilamide, sulfamethoxazole and sulfamethoxypyridazine, each of the four compounds was run

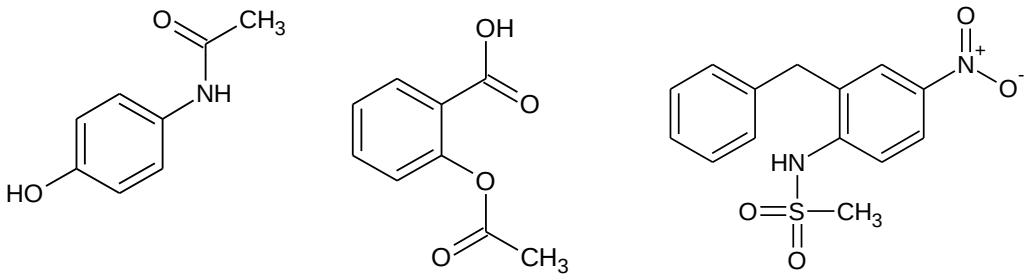
on the CCC separately allowing the identification of the elution order. Sulfanilamide, sulfapyridine and sulfamethoxazole were found to have elution times of 11.9, 12.8 and 12.8 minutes respectively, with sulfamethoxypyridazine having an elution time of 15.8 minutes. This suggested that the first peak in the CCC chromatogram contained three compounds with the second peak being sulfamethoxypyridazine. This agreed with the K_d values obtained using shake flask HPLC. The measured K_d value for sulfamethoxypyridazine was 0.8 calculated using CCC which matched the prediction exactly, with the predicted K_d value for sulfanilamide being 0.2 from the CCC K_d value. The predictions for sulfamethoxazole and sulfapyridine were less accurate being 0.7 and 1.2 units away from the prediction. Given the structural similarities between the four molecules, it is unclear as to why the predictions for these two molecules are poor.

Table 45 - A comparison between the experimentally determined K_d values from the CCC chromatogram and the predicted K_d values for sulfanilamide, sulfamethoxypyridazine, sulfamethoxazole and sulfapyridine in HEMWat 14.

Compounds	Predicted K_d values from the PLS QSAR model	Experimental K_d value obtained using retention time on CCC chromatogram	Experimental K_d value obtained using shake flask HPLC method
Sulfanilamide	0.3	0.5	0.1
Sulfamethoxypyridazine	0.8	0.8	1.7
Sulfamethoxazole	1.2	0.5	NR
Sulfapyridine	1.7	0.5	0.1

5.3.1.3. Separating acetaminophen, acetylsalicylic acid and nimesulide

The next mixture consisted of acetaminophen, acetylsalicylic acid and nimesulide (Figure 48) with their pKa values being 9.38 (Drugbank, 2015), 3.49 (Drugbank, 2015) and 6.5 (Rainsford, 2005) respectively. This mixture was selected due to the large range in the pKa values of the compounds.



A B C

Figure 48 - The molecular structures of acetaminophen (A), acetylsalicylic acid (B) and nimesulide (C).

The same method used in section 5.3.1.1 was employed to select a HEMWat system for a trial separation. However, to increase the range of the pKa values being tested, acetylsalicylic acid was added, despite it having been used in the training set. This also allowed a comparison between the shake flask HPLC and the CCC determined K_d values due to the discrepancy previously encountered (section 3.7.2).

Table 46 – The predicted K_d values for acetaminophen and nimesulide from the six QSAR models generated using PLS for each of the six HEMWat systems.

Compounds	Predicted K_d values from the six QSAR models for six HEMWat systems					
	8	14	17	20	22	26
Acetaminophen	3.6	0.4	0.0	0.0	0.0	0.0
Nimesulide	128.8	25.1	0.6	0.3	0.1	0.1

Table 47 - The experimental determined K_d values for acetylsalicylic acid using HPLC. NR indicates that the K_d values were so extreme that they could not be recorded i.e. one of the HPLC peaks was not large enough to give a signal-to-noise greater than five or they did not meet reproducibility criteria a %RSD of 10% of the triplicates.

Compounds	Experimentally determined K_d values from the six QSAR model for six HEMWat systems					
	8	14	17	20	22	26
Acetylsalicylic acid	19.85	1.63	0.19	0.05	NR	NR

HEMWat 14 was chosen as the system that was the most likely to achieve the separation of the mixture from the predicted K_d values (Table 46) as it was the only system that would avoid co-elution whilst maintaining K_d values below 2 for the majority of the compounds. The predictions and K_d values obtained by HPLC suggested that acetaminophen and acetylsalicylic acid should be separated in that order with nimesulide retained in the column. The displaced volume of the equilibrated system was 12 ml, therefore the stationary phase retention was 40%.

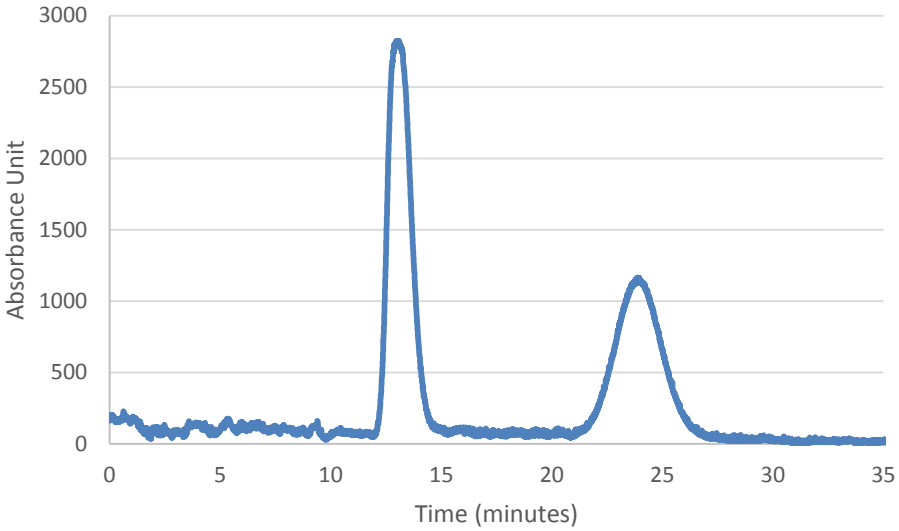


Figure 49 – The mixture of acetaminophen, acetylsalicylic acid, and nimesulide was separated using HEMWat 14 prepared by mass and left to equilibrate overnight. The acetylsalicylic acid eluted at 13 minutes, the acetaminophen eluted at 24 minutes with the nimesulide retained on the column so did not result in a peak.

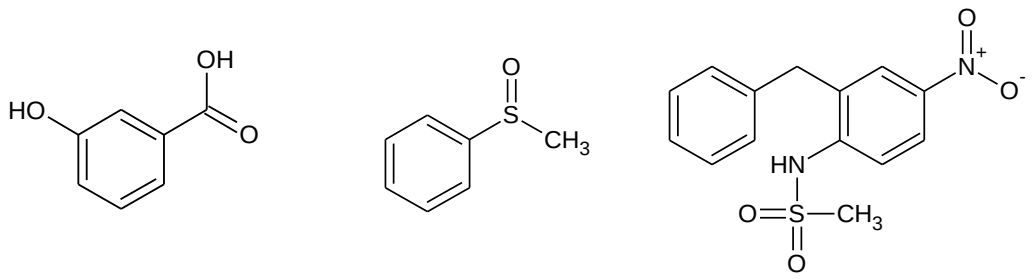
Initially, it was thought that the chromatogram was as predicted (Figure 49). However, when the K_d values obtained by HPLC and CCC were compared, a different elution order was suggested (Table 48). To conclusively prove the identity of the peaks, the spectra were examined. The first peak present in the chromatogram contained one λ_{max} matching the acetylsalicylic acid spectrum whilst the second peak had two λ_{max} points corresponding to the acetaminophen. The predicted K_d value for acetaminophen is only 0.3 units away from that measured by HPLC (Table 48). However, it is 0.7 units away from the K_d value determined using CCC. This may be due to these molecules being ionisable as there had been a discrepancy observed been the K_d values obtained on the CCC and using the shake flask HPLC method (section 3.7.2).

Table 48 – Comparison of experimentally determined K_d values from the CCC chromatogram and the predicted K_d values of acetaminophen, acetylsalicylic acid and nimesulide in HEMWat 14.

Compounds	Predicted K_d values from the PLS QSAR model	Experimental K_d value obtained using retention time on CCC chromatogram	Experimental K_d value obtained using shake flask HPLC method
Acetaminophen	0.4	1.1	0.1
Acetylsalicylic acid	-	0.4	1.6
Nimesulide	25.1	Large	34.5

5.3.1.4. Separating 3-hydroxybenzoic acid, methyl phenyl sulfoxide and nimesulide.

The next test mixture consisted of 3-hydroxybenzoic acid, methyl phenyl sulfoxide and nimesulide (Figure 50). The pKa values of 3-hydroxybenzoic acid are 4.08 (Chemicalbook, 2015) and 6.5 (Rainsford, 2005) respectively. This mixture was chosen due to their similarity of a benzene ring being present in each molecule but with range of functional groups.



A B C

Figure 50 – The molecular structures of 3-hydroxybenzoic acid (A), methyl phenyl sulfoxide (B) and nimesulide (C).

The same method used in section 5.3.1.1 was employed to select a HEMWat system for a trial separation. The addition of 3-hydroxybenzoic acid was again due to the discrepancy observed between the HPLC and CCC determined K_d values of ionisable compounds (section 3.7.2).

Table 49 - The predicted K_d values for 3-hydroxybenzoic acid, methyl phenyl sulfoxide and nimesulide from the six QSAR models generated using PLS for each of the six HEMWat systems.

Compounds	Predicted K_d values from the six QSAR models for six HEMWat systems					
	8	14	17	20	22	26
Methyl phenyl sulfoxide	3.4	0.4	0.0	0.0	0.0	0.0
Nimesulide	128.8	25.2	0.6	0.3	0.1	0.1

Table 50 - The experimental determined K_d values for ibuprofen using HPLC. NR indicates that the K_d values were so extreme that they could not be recorded i.e. one of the HPLC peaks was not large enough to give a signal-to-noise greater than five or they did not meet reproducibility criteria a %RSD of 10% of the triplicates.

Compounds	Experimentally determined K_d values from the six QSAR model for six HEMWat systems					
	8	14	17	20	22	26
3-Hydroxybenzoic acid	10.77	0.71	0.05	0.03	NR	0.00

HEMWat 14 was nominated as the system that was the most likely to achieve the separation of the mixture from the predicted K_d values (Table 49). This system was chosen due to the very low K_d values predicted in systems 17, 20, 22 and 26 suggesting that the compounds would coelute and the K_d values of the three compounds all being above two in HEMWat 8. The displaced volume of the equilibrated system was 12 ml, therefore percentage stationary phase retention was 40%.

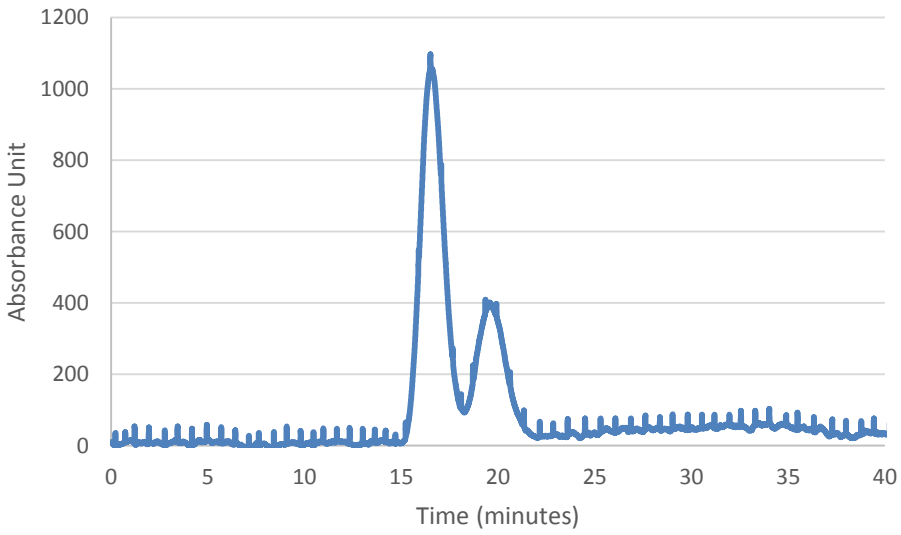


Figure 51 - The separation methyl phenyl sulfoxide was separated using HEMWat 14 prepared by mass and left to equilibrate overnight. The methyl phenyl sulfoxide eluted at 18 minutes, the 3-hydroxybenzoic acid at 21 minutes with the nimesulide retained on the column so did not result in a peak.

As can be seen from Figure 51, two peaks were observed in the chromatogram. However, baseline separation was not achieved. The identities of the peaks were confirmed by comparing the K_d values obtained using the shake flask HPLC method. From Table 37, the HPLC determined K_d values suggested that the methyl phenyl sulfoxide eluted later than has been predicted with the 3-hydroxybenzoic acid eluting as predicted.

Table 51 - Comparison of experimentally determined K_d values from the CCC chromatogram and the predicted K_d values of 3-hydroxybenzoic acid, methyl phenyl sulfoxide and nimesulide in HEMWat 14.

Compounds	Predicted K_d values from the PLS QSAR model	Experimental K_d value obtained using retention time on CCC chromatogram	Experimental K_d value obtained using shake flask HPLC method
Methyl phenyl sulfoxide	0.4	0.9 or 1.2	1.7
3-Hydroxybenzoic acid	1.2		0.7
Nimesulide	25.1	Large	34.5

The spectrum of first peak in the CCC chromatogram matched the spectrum observed for methyl phenyl sulfoxide with two λ_{max} values. The second peak corresponds to 3-hydroxybenzoic acid. This suggests that the elution order observed in the CCC was

not the elution order suggested by the HPLC readings. This is likely to reflect the ionisable nature of the 3-hydroxybenzoic acid and the difficulty of K_d measurement of ionising molecules (section 3.7.2)

5.3.2. Conclusion

The HEMWat systems selected on the basis of the predicted K_d values for separating the four test mixtures achieved separation in three of the cases. In the cases where the predictions were poor, the compounds tended to be ionisable. As the training set of compounds used to build the model contained neutral or neutralised acids and bases compounds, this is not unexpected. With a larger, more diverse training set of compounds, including acids and bases, the accuracy of the predictions from the model is likely to improve. However, as a proof of concept these results demonstrate that applying QSAR models to the prediction of HEMWat systems for use in CCC is promising.

5.4. Increasing the practical use of the model

The predictive ability of a QSAR is wholly dependent on the accuracy of the experimental data used to train it. This need to maintain a high level of accuracy led to the measurements of the K_d values using the shake flask HPLC method, not necessarily being measured in conditions that industry or academia would find convenient to translate into an experimental separation on the CCC. For example, the solvent systems used in section 5.3, had been prepared by mass using the solvent ratios in Table 1. The systems were prepared by mass to remove the influence of temperature and due to the fact that the balance allowed measurements to four decimal places more than the pipettes and measuring cylinders, used to measure volume. Despite these advantages, making large quantities of solvent systems by mass can be difficult and does not allow for “mixing on demand”, making it wasteful. “Mixing on demand” is a method of creating solvent systems for CCC runs using a quaternary pump, such as those fitted to many commercial HPLC units. With each of the four HEMWat solvents on one of the four lines, the solvent systems can be made up in situ using the percentage composition of both phases of the solvent system. This data was obtained from Garrard’s paper (Garrard, 2005). This reduces waste whilst having the added benefit of reducing the amount of time required to carry out the separation. A further change investigated, was the removal of the pH modifier as a crude mixture may include both acids and bases. In this instance, to avoid the possible

formation of salts, it would be preferable to use unadjusted HEMWat. The QSAR models had been shown to achieve separation with the predicted system when the exact conditions of the training set were used. However, it was important to assess the extent of the change in the CCC chromatogram when the predicted system was prepared using more commonly used conditions.

5.4.1. The effect of using solvent systems prepared using "mixing on demand"

5.4.1.1. Method

Following the experimental details in 2.3.3, mixtures of compounds were initially separated on the CCC centrifuge using solvent systems that were made up by mass. The same mixtures were then separated using CCC again, but this time using solvent systems made up by "mixing on demand".

5.4.1.2. Separating uracil, phenol, o-terphenyl and triphenylene

It had previously been seen that o-terphenyl and triphenylene could be separated from each other in HEMWat 26 with uracil separated from phenol in HEMWat 14, when the systems were prepared by mass (see section 5.3.1.1). These separations were repeated with the HEMWat systems prepared by "mixing on demand". Figure 52 demonstrates the difference between the two chromatograms with the red line indicating the run using the HEMWat 26 prepared by "mixing on demand" and the blue line indicating the run using the HEMWat 26 made up by mass.

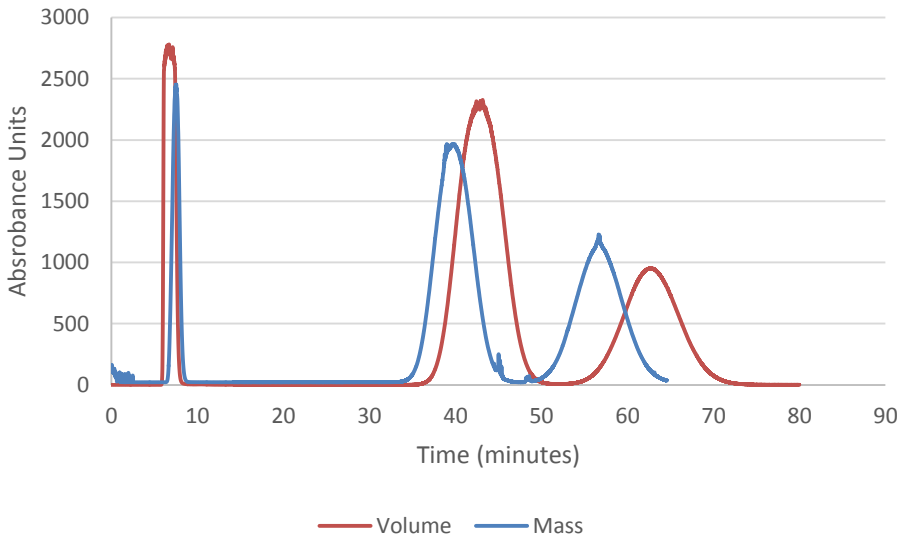


Figure 52 – The CCC chromatogram for uracil, phenol, o-terphenyl and triphenylene in HEMWat26 with 0.1% TFA in water replacing water. The chromatogram in red represents the separation carried out using HEMWat 26 that had been prepared by "mixing on demand" i.e. volume. The chromatogram in blue represents the separation carried out using HEMWat 26 made up by mass.

From Figure 52, it can be seen that uracil and phenol continue to coelute with the solvent front with separation still achieved between o-terphenyl and triphenylene. However, both compounds eluted later and therefore have a higher K_d value when using the solvent systems made using the "mixing on demand" method (Table 52).

Table 52 – The K_d values of uracil, phenol, o-terphenyl and triphenylene in HEMWat26 determined using CCC when the systems was prepared by mass and by volume ("mixing on demand").

Compounds	Experimental K_d values from HEMWat 26 prepared by volume and each phase mixed on demand	Experimental K_d values from HEMWat 26 prepared by mass and left to equilibrate overnight before the layers separated	Difference between K_d values
Uracil	0.1	0.1	0.0
Phenol			
O-terphenyl	2.6	2.4	0.2
Triphenylene	4.1	3.7	0.4

The separation of uracil and phenol in HEMWat 14 was also repeated using a solvent systems prepared using "mixing on demand".

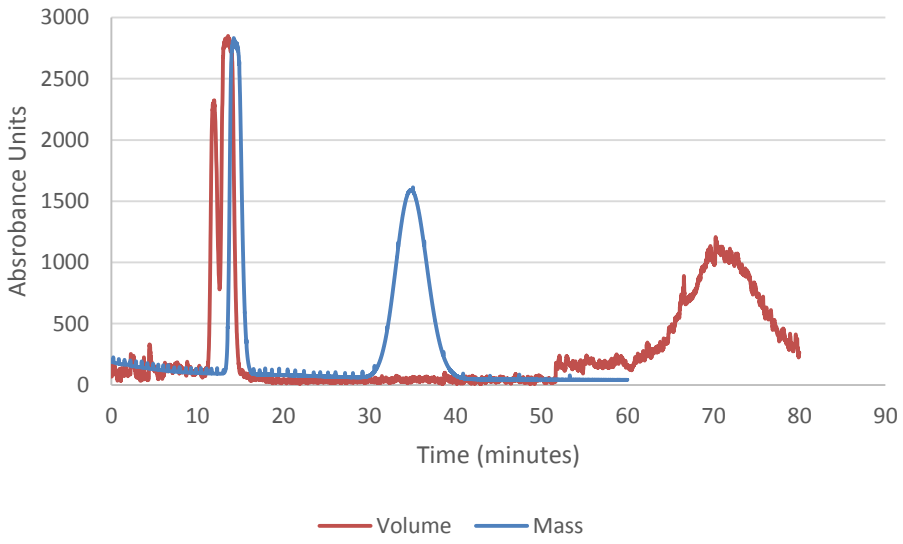


Figure 53 – The CCC chromatogram from the separation of uracil, phenol, o-terphenyl and triphenylene in HEMWat 14 prepared by mass (blue line) and volume ("mixing on demand" – red line).

As can be seen in Figure 53, the separation of uracil and phenol was once again achieved with o-terphenyl and triphenylene both retained. The uracil peak was eluted with the solvent front in both runs. However, the phenol eluted much later in the system prepared using "mixing on demand". This led to a K_d values of nearly twice that measured when the system was made up by mass (Table 53). This is a disadvantage as the run time increased from 40 minutes when the solvent systems were made up by mass, to 80 minutes when the solvent systems were made up using "mixing on demand".

Table 53 – The K_d values of uracil, phenol, o-terphenyl and triphenylene in HEMWat 14 experimentally determined using CCC from systems made by volume and mass.

Compounds	Experimental K_d values from HEMWat 26 prepared by volume and each phase mixed on demand	Experimental K_d values from HEMWat 26 prepared by mass and left to equilibrate overnight before the layers separated	Difference between K_d values
Uracil	0.3	0.6	0.3
Phenol	5.2	2.3	2.9
O-terphenyl	Large	Large	-
Triphenylene	Large	Large	-

5.4.1.3. Separating tryptamine, quinine, reserpine and lidocaine

The next example mixture that was examined for the impact of preparing the solvent systems by “mixing on demand” contained bases. Tryptamine, quinine, reserpine and lidocaine were all part of the training set so it was important to establish that they were similarity affected as the neutral and acidic compounds in the previous test mixture had been. The mixture was separated using CCC and HEMWat 17 prepared by mass and then by “mixing on demand”. The experimental details can be found in section 2.3.2 and 2.3.3.

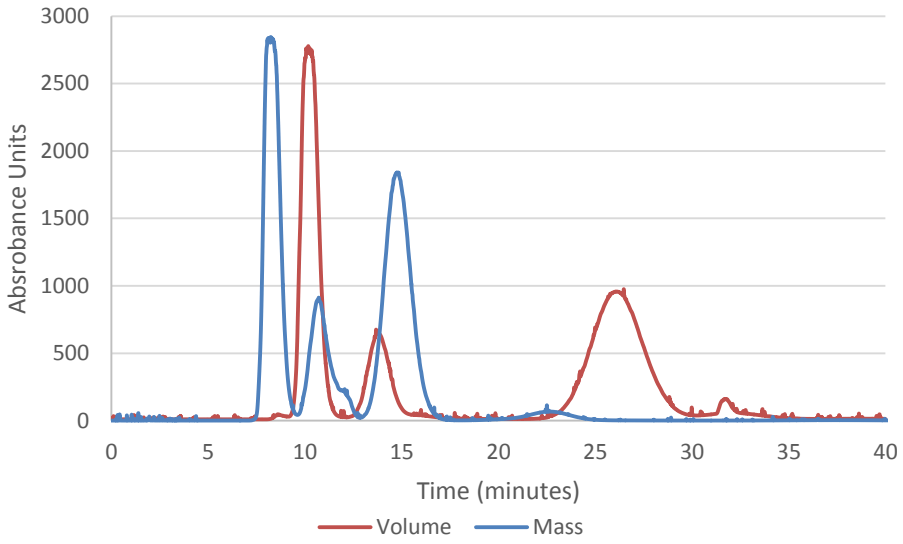


Figure 54 – The CCC chromatogram of the separation of tryptamine, quinine, reserpine and lidocaine in HEMWat 17 prepared by mass or volume (mixing in demand). The water used in both cases to prepare the HEMWat system contained 1% NH_4OH.

As can be seen in Figure 54, separation is still achieved using HEMWat 17 prepared using “mixing on demand”. As was seen with the previous example the K_d values for all of the compounds had increased (Table 54) leading to a longer run time.

Table 54 – The experimentally determined K_d values of tryptamine, quinine, reserpine and lidocaine in HEMWat 17 determined using CCC from systems made by volume and mass.

Compounds	Experimental K_d values from HEMWat 26 prepared by volume and each phase mixed on demand	Experimental K_d values from HEMWat 26 prepared by mass and left to equilibrate overnight before the layers separated	Difference between K_d values
Tryptamine	0.2	0.0	0.2
Quinine	0.5	0.2	0.3
Reserpine	0.7	0.6	0.1
Lidocaine	2.0	1.2	0.8

5.4.1.4. Conclusion

The system predicted to provide optimal separation by the models leads to separation being achieved despite the solvent system preparation method. However, all of the compounds have higher partition coefficient values when the systems are made up using the "mixing on demand" method. This leads to longer run times which is not ideal in a high pressure industrial environment. If a user wishes to use solvent systems made up by "mixing on demand", it may be worth using the HEMWat system one number below the predicted system to reduce the run time of the separation.

5.4.2. The effect of the removal of the pH modifier

5.4.2.1. Method

Following the experimental, details in section 2.3.3, the solvent systems were prepared using "mixing on demand" with pure water instead of water and 0.1% TFA in water for mixtures containing acids or 1% NH_4OH in water for mixtures containing bases.

5.4.2.1. Separating uracil, phenol, o-terphenyl and triphenylene

Acidified HEMWat 26 had been used to separate o-terphenyl and triphenylene from uracil and phenol. Both these separation were carried out in HEMWat system in which the water has been replaced with water containing 0.1% TFA. As these compounds are neutral, it was expected that the separation would be unaffected by the removal of the TFA.

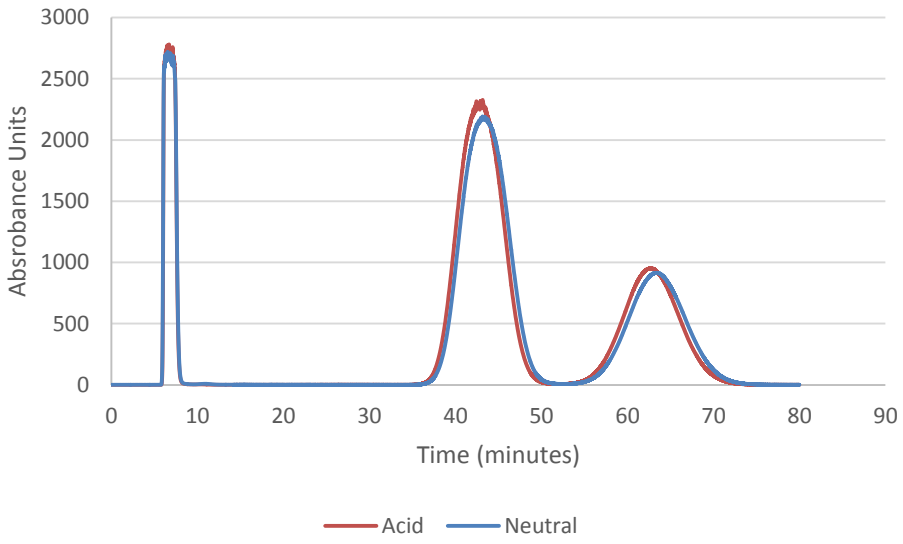


Figure 55 – The CCC chromatogram of uracil, phenol, o-terphenyl and triphenylene in HEMWat26 made by "mixing on demand". The chromatogram in red represents the CCC run using the HEMWat system prepared with water containing 0.1% TFA. The blue chromatogram represents the CCC run using HEMWat 26 prepared with pure water.

As can be seen in Figure 55, there is no change to the chromatogram meaning there is no change in K_d values. This was expected as the o-terphenyl and triphenylene are made up exclusively of benzene rings so there are no functional groups to be affected by the presence of a pH modifier. The uracil and phenol co-elute with the solvent front which will not be affected by the pH modifier.

5.4.2.2. Separating caffeine, ibuprofen and mefenamic acid

The impact of the removal of the TFA on acids, ibuprofen and mefenamic acid was also investigated. Caffeine is a neutral molecule but was used as the solvent front marker in HEMWat 22.

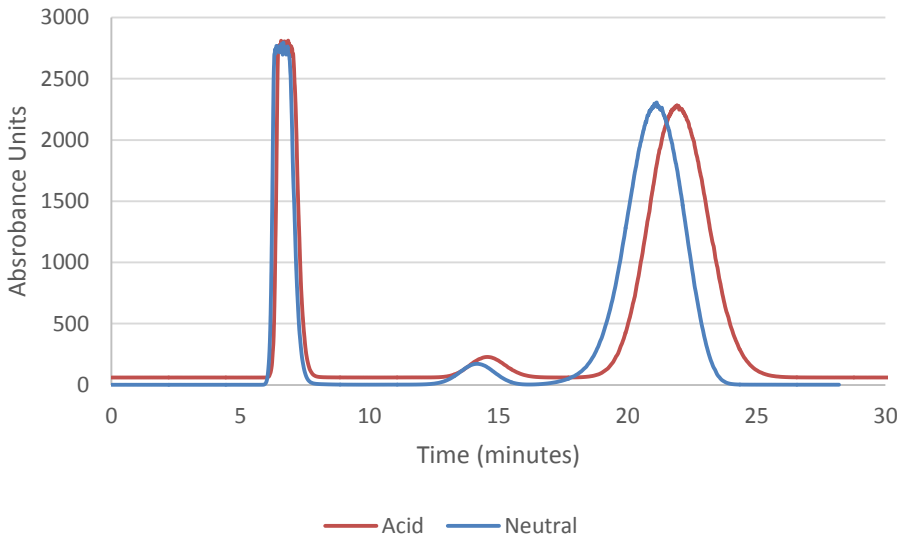


Figure 56 – The separation of caffeine, ibuprofen, mefenamic acid in HEMWat 22 prepared by "mixing on demand". The chromatogram in red represents the CCC run using the HEMWat system prepared with water containing 0.1% TFA. The blue chromatogram represents the CCC run using HEMWat 22 prepared with pure water.

As can be seen in Figure 56, there is very little difference between the chromatograms despite the removal of the pH modifier in one run. From Table 55, it can be seen that the K_d values for each compound obtained in acidified and unadjusted HEMWat differ by 0.1 for both ibuprofen and mefenamic acid with the caffeine eluting with the solvent front in both cases.

Table 55 - The experimentally determined K_d values of caffeine, ibuprofen, mefenamic acid in HEMWat 22 determined using CCC from systems prepared by "mixing on demand" with and without pH modifier. The acidified HEMWat was prepared using water containing 0.1% TFA in water with the unadjusted HEMWat being prepared with water.

Compounds	Experimental K_d values from acidified HEMWat	Experimental K_d values from unadjusted HEMWat	Difference between K_d values
Caffeine	0.0	0.0	0.0
Ibuprofen	0.7	0.6	0.1
Mefenamic acid	1.2	1.1	0.1

5.4.2.3. Separating tryptamine, quinine, reserpine and lidocaine

To assess the effect on bases, a mixture containing tryptamine, quinine, reserpine and lidocaine was run in basified and unadjusted HEMWat.

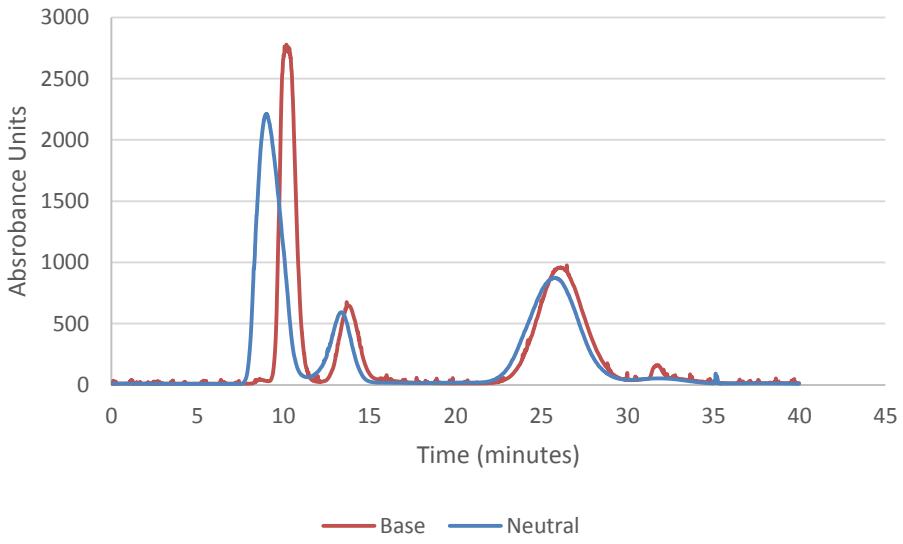


Figure 57 – The separation of tryptamine, quinine, reserpine and lidocaine in HEMWat 17 prepared by "mixing on demand". The chromatogram in red represents the CCC run using the HEMWat system prepared with water containing 1% NH_4OH. The blue chromatogram represents the CCC run using HEMWat 22 prepared with pure water.

From Figure 57, it can be seen that separation is still achieved using the unadjusted HEMWat. As can be seen in Table 56, the largest difference in K_d value was 0.1 for tryptamine and reserpine. The K_d values for quinine and lidocaine were the same as those obtained using basified HEMWat.

Table 56 – The K_d values of tryptamine, quinine, reserpine and lidocaine in HEMWat 17 prepared by "mixing on demand" with or without pH modifier, determined using CCC. The basified HEMWat was prepared using water containing 1% NH4OH with the unadjusted HEMWat being prepared using pure water.

Compounds	Experimental K_d values from basified HEMWat	Experimental K_d values from unadjusted HEMWat	Difference between K_d values
Tryptamine	0.2	0.1	0.1
Quinine	0.5	0.5	0.0
Reserpine	1.6	1.5	0.1
Lidocaine	2.0	2.0	0.0

5.4.2.4. Conclusion

The removal of the pH modifier had no impact on neutral compounds. For acid and bases the maximum change in K_d values is 0.1 which is not enough to cause co-elution or an increased run time.

6. Increasing the Appeal of the Model

The PLS models used to predict the optimal solvent systems for performing the separations in chapter 5 were generated using Simca. One of the aims of this thesis is to produce a model that has a broad appeal and does not require a large capital investment. It was thought that this could be achieved by exploiting the linear relationship that the QSAR models are based on. Therefore, this type of equation can easily be run in Microsoft Excel. As a result, the model for method development will be accessible to a large number of people, which will lead to a much wider uptake of the technology.

6.1. Transfer of the QSAR models generated using PLS from Simca to Excel

Simca calculates predicted K_d values using the scores and loadings of the model (Equation 59).

$$Y_{scaled} = tPS \times c$$

Equation 59 – Simca calculates the normalised predicted K_d value (Y_{scaled}) using the scores (tPS) of the prediction set and the loadings (c).

The score (tPS) demonstrates correlations between observations whilst the loading (c) of a model describes the correlation of a PLS component with the original variables. The closer the loading is to 1 or -1, the more significant the component is in modelling the relationship, whilst a value of 0 demonstrates that the component had no influence. The predicted K_d value (Y_{scaled}) is a weighted prediction allowing for the normalised comparison of variables and observations. To allow an unscaled prediction, the weight and offset of the normalisation must be taken into account (Equation 60).

$$Y_{unscaled} = \frac{Y_{scaled}}{Y_{ws}} + Y_{avg}$$

Equation 60 – Calculating the actual predicted K_d value ($Y_{unscaled}$) from the normalised prediction (Y_{scaled}) using Y_{avg} as the offset and Y_{ws} as the weight.

The above equation can be used to predict $\log K_d$ using the Simca software. Alternatively, QSAR models can be represented in the form of a linear equation involving the summation of the significant descriptors, multiplied by calculated

coefficients added to the residual constant. This allows the QSAR models to be transferred into Excel. The predicted K_d values calculated by the model in Excel were the same as the predicted values obtained from Simca for 22 out of 24 predictions (section 9.4). In the two cases where the Excel prediction did not match the Simca prediction, the compounds had missing values for three of the descriptors. This meant that the Excel model was treating this descriptor as zero, whereas the Simca model was calculating a non-zero number from the score and loading values. Therefore, the QSAR models produced in Simca would not work in Excel. As there were only three descriptors that had missing values for some compounds (MindistAA, MindistDA and MindistDD), they were removed and new QSAR models generated using PLS. These new models were then transferred into Excel and used for predictions, which were compared to the predicted K_d values from the original models run in Simca produced using all of the descriptors.

Table 57 – The statistical data for the training set when the predicted K_d values are plotted against the experimental K_d values and a linear line of best fit is applied. The closer the R^2 and Q^2 values are to 1 the more accurate the predictions from the model are. A R^2 value above 0.78 and a Q^2 value above 0.65 are considered acceptable.

HEMWat number	R^2		Q^2	
	New model run in Excel with three less descriptors	Original model run in Simca	New model run in Excel with three less descriptors	Original model run in Simca
8	0.75	0.69	0.58	0.66
14	0.73	0.83	0.70	0.69
17	0.70	0.81	0.70	0.65
20	0.73	0.85	0.68	0.61
22	0.69	0.89	0.69	0.80
26	0.84	0.92	0.84	0.86

From Table 57, it can be seen that the original models run in Simca had the best R^2 values for the training set for five of the six QSAR models, with the exception being HEMWat 8. Based on the R^2 values, the only new model to meet the acceptance criteria was for HEMWat 26. This suggests that the Simca models would produce the most accurate predictions. However, the Q^2 values suggested that the Excel models

had the best predictive ability in the half of systems. Interestingly, the Q^2 value for the Simca model for HEMWat 8 was higher than the Excel model. This was unexpected as the R^2 values had suggested that the QSAR model for HEMWat 8 was the only system in which the Excel model should have a greater predictive ability. It is also worth noting, that the R^2 and Q^2 values for the QSAR models for two out of the six systems differed by less than 0.02. The much smaller differences between the R^2 and Q^2 values for the Excel models can explain why in some cases the Q^2 value suggests that the Excel model performs better than the Simca model, despite the R^2 value suggesting the opposite. For three of the Excel models, the Q^2 results are the same as the R^2 results with a further two models having a difference of less than 0.05. For the Simca models, only one model had a difference between the R^2 and Q^2 values of less than 0.05 with three of the models having a difference greater than 0.1. The smaller the difference between the R^2 and Q^2 values the better, as this confirms that the good fit demonstrated by a high R^2 value translates into a good predictive ability indicated by a high Q^2 value. Therefore, the smaller differences between the R^2 and Q^2 values from the Excel models are an improvement on the Simca models. Overall, removing the three descriptors had not reduced the predictive ability of the models.

Although Q^2 can provide a good indication as to the predictive ability of a model, the most rigorous test is to externally validate the model. This was carried out using the usual four test set compounds of biphenyl, benzoquinone, tolbutamide and quinine (Figure 24). The models were used to predict the $\log K_d$ values for the compounds and compared to the experimentally determined values using R^2 and RMSE values. The latter allowed the identification of misleading R^2 values. The R^2 and RMSE values in Table 58 demonstrate that the QSAR models run in Excel produced the best predictions for five out of the six HEMWat systems. It is worth noting that the one system in which the Simca model produced the more accurate predictions was HEMWat 8. This is despite it being the only system in which the training set statistics suggested that the Excel model would provide better predictions. The R^2 and RMSE values for HEMWat 17, 20 22 and 26 differed by less than 0.1. These results indicate that the models could be moved into Excel with a minimal loss in predictive ability.

Table 58 – The R^2 and RMSE values for the prediction of the four test set compounds when the predicted values are plotted against the experimental values and a linear line of best fit is applied. The closer the R^2 values are to 1 the more accurate the predictions from the model are. A R^2 value above 0.78 and a RMSE value below 0.5 are considered acceptable.

HEMWat number	R^2		RMSE	
	New model run in Excel with three less descriptors	Original model run in Simca	New model run in Excel with three less descriptors	Original model run in Simca
8	0.66	0.85	0.80	0.67
14	0.87	0.67	0.37	0.60
17	0.85	0.82	0.41	0.43
20	0.92	0.83	0.37	0.44
22	0.97	0.94	0.20	0.27
26	0.83	0.82	0.44	0.47

6.2. QSAR models produced using PLS with only freeware and manually calculated descriptors

Having successfully demonstrated that the model could be adapted to be run in Excel, there is still the disadvantage that the would-be user must generate the descriptors for the model using specialist software. Using descriptors that are easily calculated using free software, or calculated without the need for software, will enhance the appeal of the QSAR models. Applying this criterion to the 201 descriptors reduced the number to 51 (section 9.5). The only descriptor from the "lipophilicity" category retained in the descriptor set was the ACDlogP term as this can be obtained using the ACD freeware (ACDlabs, 2015). From the "hydrogen bonding" descriptors any that can be manually calculated were kept, e.g. the number of hydrogen bond acceptors, but others were removed, e.g. the solvent accessible surface hydrogen bond acceptor area. The descriptors that fall into the category of "size/shape", where subject to the same criterion leading to descriptors that could be manually calculated, e.g. Rotbond, being included. No descriptors from the categories of "Charge/polarity", "Topology" and "Drugability" were included with every descriptor from in the category "Atom Counts" retained.

These descriptors were used to generate PLS models using the method described in section 2.2.1. The performance of each model was assessed using the R^2 and Q^2

values for the training set. For HEMWat systems 8, 14, 22 and 26, the best performing models were obtained after removing the descriptors with a VIP value of less than one, once. For HEMWat systems 17 and 20, the best performing models were obtained after removing the descriptors with a VIP value of less than one, twice.

Table 59 – Statistics from the PLS models from only freeware and manually calculable descriptors.

HEMWat number	R^2 training set	Q^2 training set	R^2 test set	RMSE test set
8	0.79	0.62	0.78	0.91
14	0.81	0.69	0.82	0.54
17	0.82	0.66	0.99	0.32
20	0.59	0.54	0.75	0.52
22	0.90	0.77	0.97	0.45
26	0.83	0.76	0.77	0.50

Applying the above approach, the generated QSAR models for five of the six systems, have R^2 values for the training set above 0.78 and four of the six have Q^2 values above 0.65 (Table 59). The R^2 values for the training set for the HEMWat 8, 17 and 22 models are higher than either of the values obtained from any previous models. This indicates that the models are capable of producing good predictions. However, for the HEMWat 20 model, both the R^2 and Q^2 values for the training set are below the acceptance criteria, which suggests it will not be able to produce such accurate predictions. To fully test the models, they were externally validated using the usual four test compounds of benzoquinone, biphenyl, tolbutamide and quinine (Figure 24). When the QSAR model for HEMWat 8 was externally validated, the predictions had a considerably higher RMSE value when compared to the other five models, due to the model having a Q^2 value of 0.62 which is below the acceptance criteria. This suggests that the R^2 value from the test set is misleading. This may be due to predictions clustering around two points resulting in a high R^2 value, despite poor predictions. For the HEMWat 17 and 22 models the R^2 values for the test set were higher than the values obtained from any previous models. For four out of the six models, the R^2 values for the test set were above 0.75. However, from the RMSE data is can be seen that only three of the six models meet the acceptance criteria with a RMSE value of less than 0.5. Two of the models (HEMWat 14 and 20) do not reach the criteria as

their RMSE values are 0.52 and 0.54 which are only just above the acceptance criteria of 0.5.

To further examine why the R^2 values from the test set were misleading, the differences between the experimentally determined $logK_d$ values and the predicted $logK_d$ values were investigated.

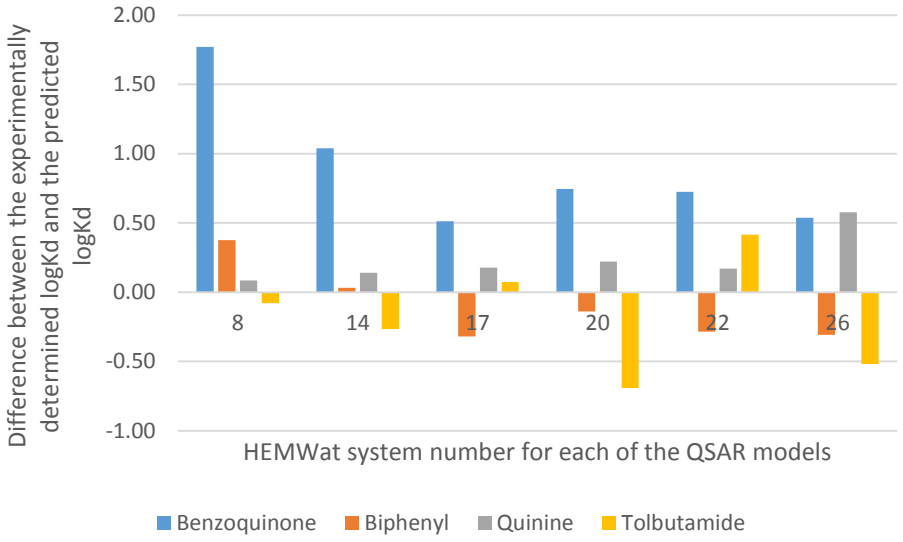
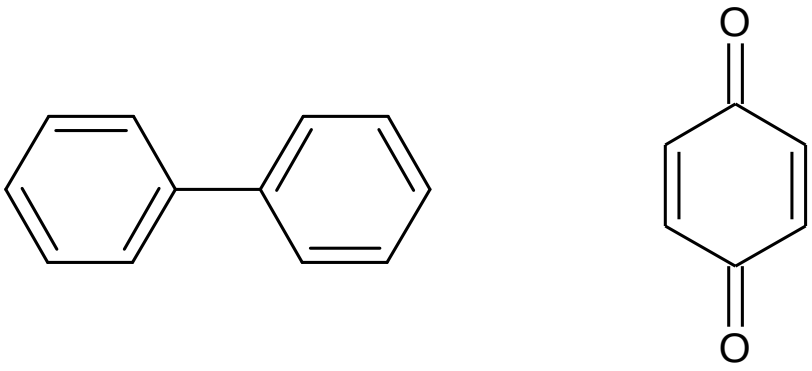


Figure 58 – The difference between the experimentally determined $logK_d$ values (see section 3.7.5 for the experimental procedure used to determine these values) and the predicted $logK_d$ values for the four test set compounds obtained from each of the six QSAR models generated using PLS, one for each HEMWat system.

Figure 58 shows the difference between the experimental and predicted K_d values for the four test set compounds. The predictions for benzoquinone in HEMWat 8 and 14 are very poor with difference of 1.77 and 1.04 respectively, with neither of these models predicting the $logK_d$ value within one log unit of the experimental value. However, none of the six models predicted benzoquinone's $logK_d$ value within the acceptance criteria of 0.5 log units of the experimental value. In contrast, each of the six models predicted the $logK_d$ values for biphenyl within the acceptance criterion of 0.5 log units. The predictions of the $logK_d$ values for tolbutamide and quinine from the model for HEMWat 26, were just outside of the acceptance criterion with a difference of -0.52 and 0.58 respectively. The other five models predicted the $logK_d$ values for quinine to within 0.5 log units with four models meeting this criterion for tolbutamide. It is likely that the biphenyl was well predicted due to it comprising of benzene rings

only with no additional functionality (Figure 59). There are compounds in the training set that are also purely made of benzene rings (e.g. naphthalene) so it is foreseeable that the model is able to produce such accurate predictions for similar molecules. On the other hand, the benzoquinone is made up of a six membered carbon ring with two carbonyl groups and two C=C groups (Figure 59). This prevents the compound from achieving aromaticity. This type of functionality is not well represented in the training set, so the models have been unable to predict the $\log K_d$ values accurately.



(A) (B)

Figure 59 - The molecular structure of biphenyl (A) and benzoquinone (B)

The models produced using PLS and 51 descriptors are able to provide reasonable predictions. The exception is the QSAR model for HEMWat 8 which is extremely poor. However, by exploiting the linear relationship between $\log K_d$ and HEMWat system number the removal of one QSAR model should not have a large impact of the predicted ability of the overall model.

As can be seen in Table 60, the difference between the experimentally determined and the predicted $\log K_d$ values for the test compounds in HEMWat 8 is less than 0.15 for three out of the four compounds. The predicted values were obtained by extrapolating the linear line of best fit from the experimentally determined $\log K_d$ values in HEMWat 14, 17, 20, 22 and 26. This will still give a good starting point for users who cannot invest in specialist software e.g. C-Lab (description generator) or Simca software (prediction generator).

Table 60 – A comparison of the experimentally determined and predicted logK_d value for the four test compounds in HEMWat 8. The predicted logK_d values were obtained from the extrapolation of the linear line of best fit from plotting the experimentally logK_d values in HEMWat 14, 17, 20, 22 and 26.

Compounds	Experimentally determined logK_d values	Predicted logK_d values	Difference
Benzoquinone	-0.67	-0.53	0.14
Biphenyl	2.41	2.44	0.03
Quinine	1.51	1.43	0.08
Tolbutamide	2.24	1.71	0.53

6.3. Conclusion

It has been demonstrated that the PLS QSAR models run in Excel will give predictions in line with the PLS QSAR models from Simca. Since the use of this model requires access to the C-lab software, which is not always available, Excel models were developed with only the descriptors obtainable from freeware and those which can be manually calculated. It has been shown that the Excel models will be able to provide predictions that are as good as the Simca model.

7. Conclusion, Further Research and Final Comments

7.1. Conclusion

One of the hurdles preventing the uptake of CCC and CPC by industry is the lack of an established protocol for determining the solvent system to be employed. A full and comprehensive review of current literature was carried out to establish the current situation. The many, varied, practical solvent selection methods were reviewed and it was concluded that each method suffered from the disadvantage of requiring time-consuming experiments with potentially expensive or scarce compound. Solvent selection using a computational model would overcome this disadvantage. However, the models currently available often require expensive software and are time-consuming. QSAR models have the potential to offer an alternative solution. This resulted in the following aims: establishing a protocol for the accurate and reproducible experimental measurements of partition coefficient (K_d) values with a second aim of using this data to develop QSAR models for predicting the optimal HEMWat solvent system for the separation of compounds in CCC.

To achieve this, the physical factors which affected the reproducibility of the experimental measurement of K_d were investigated. It was found that it was important to control the temperature and equilibration time of the solvent systems and not to use a co-solvent for the initial dissolution of the compounds as they lead to large variations in the measured K_d values. The concentration of the initial sample was determined not to impact the experimental K_d measurement as the variation between the K_d values obtained using a saturated solution and a solution five times more diluted were within experimental error. By using this protocol reproducible K_d measurements were possible for neutral compounds with six experimentally determined K_d values having a %RSD of less than 10. Unexpectedly, it was found that reproducible results were much harder to achieve for ionisable compounds with the highest %RSD being 123.40% for ibuprofen in HEMWat 20. Despite a consistently low standard deviation of less than 0.1 across all experimental measurements, the %RSD was above the acceptance criteria of 10% in the systems in which the compounds had a K_d value of 1. Although this was likely to be due to the small size of the $\log K_d$ value, since this was the value that the models were trying to predict, the accuracy was especially important. This variation in the $\log K_d$ values of acidic compounds was minimised with the addition

of a small amount of pH modifier to neutralise the compounds leading to reproducible results when measured by HPLC. It was therefore decided to measure the $\log K_d$ values of acids in acidified HEMWat (0.1% TFA added to the water of the HEMWat system). In addition, it was decided to measure the K_d value of bases in basified HEMWat (1% NH_4OH added to the water of the HEMWat system). A smaller level of variation had been seen with the $\log K_d$ values of basic compounds but it was decided to add pH modifier to allow the pH to be controlled. As pH was known to affect K_d values for ionisable compounds, this control increased the likelihood of reproducible results. This was demonstrated by a decrease in standard deviation in each case where pH modifier was added. The fact that using the standard methodology produced K_d measurements that were reproducible 3 months apart, in different laboratories, with many different pieces of equipment, lead to the conclusion that the first aim of this work had been met.

In addition to this standard methodology for obtaining K_d values by HPLC, CCC can also be used to obtain K_d values. The same value should be determined by both methods. However, there was found to be discrepancy between the K_d values from the HPLC and CCC for carboxylic acids that was not fully explained. A change in the solvent system composition due to the hydrolysis of ethyl acetate was ruled out as neutral compounds were unaffected. Additionally, an effect from the Teflon tubing retaining acid molecules was also dismissed due to there being no difference in K_d values measured after runs containing acid or after overnight soaking in methanol. It has been hypothesised that the higher K_d values obtained by CCC may be due to a change in pH during the CCC run. It was not possible to conclusively prove or disprove this due to the challenges associated with measuring pH in organic solvents.

The literature review highlighted the need for a training set of diverse compounds to build the QSAR models. This was achieved by examining the five Abraham's parameters (hydrogen bonding acidity, A, hydrogen bonding basicity, B, polarisability/polarity, S, excessive molar refraction, E, and McGowan volume, V) for a variety of compounds, which were added and removed from the training set until a selection of compounds covering a large range of each of the Abraham's parameters was found. The diversity of the training set was confirmed by the large spread compounds across parameter space found by PCA. The $\log K_d$ values for each of the

training set compounds were measured using the standard methodology developed in chapter 3.

The second aim of this thesis was to use the $\log K_d$ data to develop QSAR models for predicting the optimal solvent system for the separation of compounds in the HEMWat systems. Once this data had been collected using the standard methodology, four mathematical methods were used to generate QSAR models. Two were machine learning techniques, random forest (RF) and support vector machines (SVM) with two multivariate analysis techniques, multiple linear regression (MLR) and partial least squares (PLS) also examined. Interestingly, the predictive ability of the models generated using RF and SVM, provided poor predictions with no model meeting the acceptance criteria for the R^2 and RMSE values of the training set. However, it was established that this was likely to be due to restrictions of the AutoQSAR platform, not the methods themselves due to the PLS models produced by this platform also being very poor. With software packages specific to these machine learning techniques, there is potential that they could still be used to produce accurate models. These techniques will also benefit from a larger data set.

The two regression techniques used to develop the QSAR models were PLS and MLR. The QSAR models generated using MLR and PLS were more promising when tested against the randomly selected four test set compounds. It was decided to use the models generated using PLS as this method is not affected by collinearity and it less likely to be overfitted than the models produced using MLR.

Having developed QSAR models, these were externally validated to assess whether the second aim of this thesis has been achieved. The accuracy of the selected QSAR models predictions were tested using two different methods. The first was to identify literature examples in which HEMWat had been used to perform successful separations, which was then compared to the system predicted by the model. This method allowed the model to be tested against separations that were known to be successful. It was found that the predictions for neutral compounds were within three HEMWat system numbers of the system used to perform the successful separation. For the neutralised compound, tiamulin which was run in the presence of 1% NH_4OH, the models also predicted the system to within one HEMWat system number. As the data used to train the model had only contained neutral or neutralised compounds, it

was expected that the predictions for the strongly ionisable compounds were poor. Linking the QSAR models to the logD curves of compounds may overcome this and would be interesting further work. However, it could also be due to the discrepancy between the HPLC and CCC K_d values for ionisable compounds, discussed above.

The second validation method was to predict the HEMWat system in which a number of compounds would be successfully separated. The K_d values of all of the compounds in the six systems were predicted and the system that predicted K_d values with a separation factor larger than 1.5 was selected. The predicted HEMWat system was then used to perform a CCC run to see if the separation was successful. It was found that separation was achieved in the three out of four cases, whilst the K_d values for ionisable compound were often twice as large as predicted. It was also found that separation could be achieved using the predicted system made up by mixing on demand as opposed to mass. However, the separation required a longer run time which is a disadvantage. It was found that the removal of pH modifier only had a small impact on K_d value with a change of 0.1 observed.

Finally, the usability of the QSAR models was enhanced by transferring the models into Excel, which is widely available compared to SIMCA. This was successfully achieved. The descriptor set was then refined to only include descriptors that could be calculated manually or using freeware e.g. ACDlabs. When new QSAR models were generated using this refined descriptor set, the training set met the acceptance criteria for R^2 in five out of six cases and Q^2 in four out of six cases suggesting that the models were capable of producing accurate predictions.

Overall, the developed models were able to predict a system that lead to the successful separation of neutral and neutralised compounds. Further understanding is required to produce a model that provides accurate predictions of the K_d values of ionisable compounds. Having shown that the models could be transferred into Excel without a loss of predictive ability and developed models that only require manually calculated or freeware descriptors, these models can provide a user with a predicted solvent system for a CCC/CPC run which is likely to result in successful separation.

7.2. Future Research

This work has shown that QSAR models can be used to select HEMWat systems to separate groups of compounds without the need for experiments or expensive

software. However, the QSAR models produced are limited to predicting a solvent system from the HEMWat table (Table 1). The HEMWat system may not always be appropriate for all compounds, so generating QSAR models for other solvent system families, would be the next step. This may be achieved using the physiochemical properties of solvents. For example, a recent paper by Lesellier set out a spider diagram approach for representing the physiochemical properties of solvents (Lesellier, 2015). This demonstrates the potential of modelling to change solvent systems in cases where HEMWat does not provide high enough solubility for preparative loading or causes instability/denaturation of compounds. Liquid handling robots could be used to quickly generate K_d data in many other HEMWat systems. However, the standard methodology may need to be adapted for use with a robot to take into account the volatility of the HEMWat solvents. Also, the future training sets should include compounds that occupy any areas of parameter space that are currently empty, for example, compounds containing -CF_3 groups should be included. The training set would also benefit from the addition of more neutralised acids and bases. Furthermore, an improvement in the model would come from its ability to predict the K_d values of partially ionised compounds. This could be done using the pKa curve for a compound.

Once the data set of logK_d values has been doubled, the QSAR models could be regenerated using RF and SVM in specialist software that contains some manual functionality. This will increase the probability of producing models with a better predictive ability. For the models generated using RF, the ability to optimise the number of trees and sub trees may lead to improved models. The software used to generate the QSAR models by SVM should allow different Kernel functions to be applied to the data set as a function, other than the radial basis function, may provide a better fit for the data. It is also likely that both methods would benefit from the ability to select the most significant descriptors to be included in the model.

7.3. Final Comments

The application of QSAR models to the solvent selection for optimal separation by CCC has been shown to be a promising technique in this proof of concept work. It is hoped that this will improve the technique's appeal to industry and aid the establishment of CCC and CPC as one of the predominant purification techniques.

8. References

Abbott T. P. and Kleiman R. Solvent Selection Guide for Counter-Current Chromatography [Journal] // Journal of Chromatography A. - 1991. - Vol. 538. - pp. 109-118.

Abraham M. H. [et al.] The Partition of Compounds from Water and from Air into Wet and Dry Ketones [Journal]. - [s.l.] : New J. Chem, 2009. - 3 : Vol. 33. - pp. 568-573.

Abraham M. H. [et al.] Hydrogen bonding. 32. An analysis of water/octanol and water/alkane partitioning and the delta logP parameter of Seiler [Journal] // Journal of Pharmaceutical Sciences. - 1994. - 8 : Vol. 83. - pp. 1085-1100.

Abraham M. H. [et al.] Thermodynamics of solute transfer from water to hexadecane [Journal] // Journal of Chemical Society, Perkin Transactions 2. - 1990. - 2. - pp. 291-300.

Abraham M. H. and Chadha H. S. Applications of a salvation equation to drug transport properties in Lipophilicity in Drug Action and Toxicology [Journal] // VCH Weinheim. Edited by V. Pliska, B. Testa, H. Van der Waterbeemd. - 1996.

Abraham M. H. Internal Email, University College London [Journal]. - 2013.

Abraham M. H., Ibrahim A. and Zissimos A. M. Determination of sets of solute descriptors from chromatographic measurements [Journal] // Journal of Chromatography A. - 2004. - Vol. 1037. - pp. 29–47.

ACDlabs http://www.acdlabs.com/resources/freeware/chemsketch/ [Journal]. - 2015.

ACDlabs https://ilab.acdlabs.com/iLab2/ [Journal]. - 2015.

Advanced Bioprocessing Centre Brunel University Internal report [Journal].

Advanced Bioprocessing Centre Brunel University Internal Report [Journal]. - 2011.

Advanced Bioprocessing Centre Brunel University Internal Report 2012 [Journal]. - 2012.

Amat L., Carbo-Dorca R. and Ponec R. Molecular Quantum Similarity Measures as an Alternative to Log P Values in QSAR Studies [Journal] // Journal of Computational Chemistry. - 1998. - 14 : Vol. 19. - pp. 1575-1583.

Berthelot M. and Jungfleisch E. On the laws that operate for the partition of a substance between two solvents [Journal] // J. Ann. Chim. Phys.. - 1872. - 26 : Vol. 4. - pp. 396-407.

Berthod A. [et al.] Countercurrent Chromatography in Analytical Chemistry (IUPAC Technical Report) [Journal] // Pure Appl. Chem.. - 2009. - 2 : Vol. 81. - pp. 355-387.

Berthod A. and Bully M. High-speed Countercurrent Chromatography Used for Alkylbenzene Liquid-Liquid Partition Coefficient Determination [Journal] // Anal. Chem.. - 1991. - Vol. 63. - pp. 2508-2512.

Berthod A., Hassoun M. and Ruiz-Angel M. J. Alkane effect in the Arizona liquid systems used in countercurrent chromatography [Journal] // Anal Bioanal Chem. - 2005. - Vol. 383. - pp. 327–340.

Berthod A., Mallet A. I. and Bully M. Measurement of Partition Coefficients in Waterless Biphasic Liquid Systems by Countercurrent Chromatography [Journal]. - [s.l.] : Anal Chem., 1996. - 3 : Vol. 68. - pp. 431-436.

Berthod A., Ruiz-Angel M. J. and Carda-Broch S. Elution−Extrusion Countercurrent Chromatography. Use of the Liquid Nature of the Stationary Phase To Extend the Hydrophobicity Window [Journal] // Analytical Chemistry. - 2003. - 21 : Vol. 75. - pp. 5886-5894.

Blackie J. A [et al.] The Identification of Clinical Candidate SB-480848: A Potent Inhibitor of Lipoprotein-Associated Phospholipase A2 [Journal] // Bioorganic & Medicinal Chemistry Letters. - 2003. - Vol. 13. - pp. 1067-1070.

Breiman L. Bagging Predictors [Journal] // Machine Learning. - 1996. - Vol. 24. - pp. 123-140.

Breiman L. Random Forests [Journal] // Machine Learning. - 2001. - Vol. 45. - pp. 5-32.

Brown A. C. and Fraser T. R. V. On the Connection between Chemical Constitution and Physiological Action. Part. I. On the Physiological Action of the Salts of the Ammonium Bases, derived from Strychnia, Brucia, Thebaia, Codeia, Morphia, and Nicotia [Journal] // Transactions of the Royal Society of Edinburgh. - 1868. - 1 : Vol. 25. - pp. 151-203.

Caron G. and Ermondi G. Calculating Virtual logP in the alkane/water system (logPNalk) and Its Derived Parameters logPNoct-alk and logDpHalk [Journal] // Journal of Medicinal Chemistry. - 2005. - 9 : Vol. 48. - pp. 3269-3279.

Chang C.-C. and Lin C.-J. LIBSVM: a library for support vector machines. [Journal] // ACM Transactions on Intelligent Systems and Technology. - 2011. - pp. 2:27:1--27:27.

Chemicalbook
http://www.chemicalbook.com/ProductMSDSDetailCB0441823_EN.htm, 3-bromobenzoic acid [Journal]. - 2015.

Chemicalbook
http://www.chemicalbook.com/ProductMSDSDetailCB4854760_EN.htm, 3-hydroxybenzoic acid [Journal]. - 2015.

Chen L. [et al.] Rapid purification and scale-up of honokiol and magnolol using high-capacity high-speed counter-current chromatography [Journal] // Journal of Chromatography A. - 2007. - Vol. 1142. - pp. 115–122.

Chevolot L. and Foucault A. P. Counter-Current Chromatography: Instrumentation, Solvent Selection and Some Recent Applications to Natural Product Purification [Journal] // Journal of Chromatography A. - 1988. - Vol. 808. - pp. 3–22.

Colclough N. and Wenlock M. C. Interpreting physicochemical experimental data sets [Journal] // Journal of Computer-Aided Molecular Design. - 2015.

Conway W. D. An Indexing Scheme For Optimizing The Choice Of Biphasic Systems For CCC [Journal] // Journal of Liquid Chromatography & Related Technologies. - 11-12 : Vol. 24. - pp. 1555-1573.

Conway W. D. Countercurrent Chromatography: Apparatus, Theory, and Practice [Journal] // Analytical Chemistry. - 1991. - 10 : Vol. 63.

Cramer III R. D., Paterson D. E. and Bunce J. D. Comparative Molecular Field Analysis (CoMFA). I. Effect of Shape on Binding of Steroids to Carrier Proteins [Journal] // Journal of the American Chemical Society. - 1988. - Vol. 110. - pp. 5959-5967.

Csizmadia F. [et al.] Prediction of Distribution Coefficient from Structure. 1. Estimation Method [Journal] // Journal of Pharmaceutical Sciences. - 1997. - 7 : Vol. 86. - pp. 865-871.

Czermiński R., Yasri A. and Hartsough D. Use of Support Vector Machine in Pattern Classification: application to QSAR Studies [Journal] // Quantitative Structure Activity Relationships. - 2001. - Vol. 20. - pp. 227-240.

Davis A. M and Wood D. J Quantitative Structure-Activity Relationship Models That Stand the Test of Time [Journal] // Mol. Pharmaceutics. - 2013. - Vol. 10. - pp. 1183-1190.

DeAmicis C. [et al.] Comparison of preparative reversed phase liquid chromatography and Countercurrent chromatography for the kilogram scale purification of crudespinetoram insecticide [Journal] // Journal of Chromatography A. - 2011. - Vol. 1218. - pp. 6122– 6127.

Dearden J. C. and Bresnen G. M. The measurement of partition coefficient [Journal]. - [s.l.] : Quantitative Structure Activity Relationships, 1988. - 3 : Vol. 7. - pp. 133-144.

Dearden J. C., Cronin M. T. D and Kaiser K. L. E. How not to develop a quantitative structure–activity or structure–property relationship (QSAR/QSPR) [Journal] // SAR and QSAR in Environmental Research. - 2009. - 3–4 : Vol. 20. - pp. 241–266.

Dietterich T. G. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms [Journal] // Neural Computation. - 1998. - 7 : Vol. 10. - pp. 1895-1923.

Drugbank Aspirin http://www.drugbank.ca/drugs/DB00945 [Journal]. - 2015.

Drugbank http://www.drugbank.ca/drugs/DB00206, Reserpine [Journal]. - 2015.

Drugbank http://www.drugbank.ca/drugs/DB00281, Lidocaine [Journal]. - 2015.

Drugbank http://www.drugbank.ca/drugs/DB00682, Warfarin [Journal]. - 2015.

Drugbank http://www.drugbank.ca/drugs/DB01050, Ibuprofen [Journal]. - 2015.

Drugbank http://www.drugbank.ca/drugs/DB01124, Tolbutamide [Journal]. - 2015.

Drugbank http://www.drugbank.ca/drugs/DB01203, Nadolol [Journal]. - 2015.

Drugbank Paracetamol http://www.drugbank.ca/drugs/DB00316 [Journal]. - 2015.

Drugbank phenol http://www.drugbank.ca/drugs/DB03255 [Journal]. - 2015.

Drugbank Sulfanilamide http://www.drugbank.ca/drugs/DB00259 [Journal]. - 2015.

Drugbank sulfapyridine http://www.drugbank.ca/drugs/DB00891 [Journal]. - 2015.

Drugbank Uracil http://www.drugbank.ca/drugs/DB03419 [Journal]. - 2015.

Druglead Sulfamethoxypyridazine http://www.druglead.com/cds/sulfamethoxypyridazine.html [Journal]. - 2015.

Du Q.-S. [et al.] Fragment-Based Quantitative Structure–Activity Relationship (FB-QSAR) for Fragment-Based Drug Design [Journal] // Journal of Computational Chemistry. - 2 : Vol. 30. - pp. 295-304.

Dubant S. [et al.] Practical Solvent System Selection for Counter-Current Separation of Pharmaceutical Compounds [Journal] // Journal of Chromatography A. - 2008. - Vol. 1207. - pp. 190–192.

Fedotov P. S. and Khachaturov R. V. A New Approach to Describing the Regularities of Stationary Phase Retention in Countercurrent Chromatography [Journal] // Journal of Liquid Chromatography & Related Technologies . - 2000. - 5 : Vol. 23. - pp. 655-667.

Foucault A.P Chapter 4: Solvent Systems in Centrifugal Partition Chromatography [Journal] // Centrifugal Partition Chromatography, Marcel Dekker, New York. - 1995. - Vol. 68. - pp. 71-97.

Free S. M. and Wilson J. W. A Mathematical Contribution to Structure-Activity Studies [Journal] // Journal of Medicinal Chemistry. - 1964. - 4 : Vol. 7. - pp. 395-399.

Freisen J. B., Ahmed S. and Pauli G. F. Qualitative and Quantitative Evaluation of Solvent Systems for Countercurrent Separation [Journal] // Journal of Chromatography A. - 2015. - Vol. 1377. - pp. 55-63.

Friesen J. B. and Pauli G. F. Binary concepts and standardization in CCC technology [Journal] // Journal of Chromatography A. - 2009. - Vol. 216. - pp. 4237-4244.

Friesen J. B. and Pauli G. F. G.U.E.S.S. - A Generally Useful Estimate of Solvent Systems for CCC [Journal] // Journal of Liquid Chromatography & Related Technologies. - 2005. - Vol. 28. - pp. 2777–2806.

Friesen J. B. and Pauli G. F. Rational Development of Solvent System Families in Counter-Current Chromatography [Journal]. - [s.l.] : Journal of Chromatography A, 2007. - Vol. 1151. - pp. 51–59.

Fujita T. and Ban T. Structure-activity relation. 3. Structure-activity study of phenethylamines as substrates of biosynthetic enzymes of sympathetic transmitters [Journal] // Journal of Medicinal Chemistry. - 1971. - 2 : Vol. 14. - pp. 148-152.

Garrard I Internal Report [Journal]. - 2008.

Garrard I. J. Simple Approach to the Development of a CCC Solvent Selection Protocol Suitable for Automation [Journal] // Journal of Liquid Chromatography and Related Technologies. - 2005. - Vol. 28. - pp. 1923–1935.

Garrard I. J., Janaway L. and Fisher D. Minimising Solvent Usage in High Speed, High Loading, and High Resolution Isocratic Dynamic Extraction [Journal] // Journal of Liquid Chromatography & Related Technologies. - 2007. - Vol. 30. - pp. 151–163.

Ghose A. K. and Crippen G. M. Atomic Physiochemical Parameter for Three-Dimensional Structure-Directed Quantitative Structure-Activity Relationships. 1. Partition Coefficients as a Measure of Hydrophobicity [Journal] // Journal of Computational Chemistry. - 1986. - 4 : Vol. 7. - pp. 565-577.

Gramatica P. Principles of QSAR Models Validation: Internal and External [Journal] // QSAR and Combinatorial Science. - 2007. - 5 : Vol. 26. - pp. 694-701.

Hammett L. P. Linear free energy relationships in rate and equilibrium phenomena [Journal] // Transactions of the Faraday Society. - 1938. - Vol. 34. - pp. 156-165.

Han Q.-B. [et al.] A simple method to optimise the HSCCC two-phase solvent system by predicting the partition coefficient for a target compound [Journal] // Journal of Separation Science. - 2008. - Vol. 31. - pp. 1189-1194.

Hansch C. [et al.] Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients [Journal] // Nature. - 1962. - Vol. 194. - pp. 178 - 180.

Hewitson P. [et al.] Intermittent counter-current extraction as an alternative approach to purification of Chinese herbal medicine [Journal] // Journal of Chromatography A. - 2009. - Vol. 1216. - pp. 4187–4192.

HICHROM Chromatography Columns and Supplies Catalogue 9 [Journal].

Hilal S. H., Karickhoff S. W. and Carreira L. A. Prediction of the Solubility, Activity Coefficient and Liquid/Liquid Partition Coefficient of Organic Compounds [Journal]. - 2004. - 9 : Vol. 23. - pp. 709-720.

Himmel D. [et al.] Anchor Points for the Unified Bronsted Acidity Scale: The rCCC Model for the Calculation of Standard Gibbs Free Energies of Proton Solvation in Eleven Representative Liquid Media [Journal] // Chemistry - A European Journal. - 2011. - Vol. 17. - pp. 5808-5826.

Hopmann E., Arlt W. and Minceva M. Solvent system selection in counter-current chromatography using conductor-like screening model for real solvents [Journal] // Journal of Chromatography A. - 2011. - Vol. 1218. - pp. 242-250.

Hopmann E., Frey A. and Minceva M. A priori selection of the mobile and stationary phase in centrifugal partition chromatography and counter-current chromatography [Journal] // Journal of Chromatography A. - 2012. - Vol. 1238. - pp. 68-76.

Hu R. and Pan Y. Recent trends in counter-current chromatography [Journal] // Trends in Analytical Chemistry. - 2012. - Vol. 40. - pp. 15-27.

Ianciuc O. Applications of Support Vector Machines in Chemistry [Journal] // Reviews in Computational Chemistry, Vol. 23, Eds. : K. B. Lipkowitz and T. R. Cundari. Wiley-VCH, Weinheim. - 2007. - pp. 291-400.

Ignatova S. [et al.] Gradient elution in counter-current chromatography: A new layout for an old path [Journal] // Journal of Chromatography A. - 2011. - Vol. 1218. - pp. 6053– 6060.

Inoue K. [et al.] A strategy for high-speed countercurrent chromatography purification of specific antioxidants from natural products based on on-line HPLC method with radical scavenging assay [Journal] // Food Chemistry. - 2012. - Vol. 134. - pp. 2276–2282.

Ito Y. [et al.] The Coil Planet Centrifuge [Journal] // Nature. - 1966. - 5066 : Vol. 212. - pp. 985-987.

Ito Y. and Bowman R. L. Countercurrent Chromatography: Liquid-Liquid Partition Chromatography without Solid Support [Journal] // Science. - 1970. - Vol. 167. - pp. 281-283.

Ito Y. Golden Rules and Pitfalls in Selecting Optimum Conditions for High-Speed Counter-Current Chromatography [Journal] // Journal of Chromatography A. - 2005. - Vol. 1065. - pp. 145-168.

Ito Y. pH-zone-refining counter-current chromatography: Origin, mechanism, procedure and applications [Journal] // Journal of Chromatography A. - 2013. - Vol. 1271. - pp. 71-85.

Jover J., Bosque R. and Sales J. Determination of Abraham solute parameters from molecular structure [Journal] // Journal of chemical information and computer sciences. - 2004. - 3 : Vol. 44. - pp. 1098-1106.

Jozwiak K. [et al.] Use of Reversed-Phase High-Performance Liquid Chromatography in QSAR Analysis of 2,4-dihydroxythiobenzanilde Analogues [Journal] // SAR and QSAR in Environmental Research. - 1999. - Vol. 10. - pp. 509-532.

Kamlet M. J. [et al.] Linear Solvation Energy Relationships. 13. Relationship between the Hildebrand Solubility Parameter, δH, and the Solvatochromic Parameter, π* [Journal] // Jo Am. Chem. Society. - 1981. - Vol. 103. - pp. 6062-6066.

Kamlet M. J. [et al.] Solubility Properties and Biological Media. 4. Correlation of Octanol/Water Partition Coefficients with Solvatochromatic Parameters [Journal] // Journal of the American Chemical Society. - 2 : Vol. 106. - pp. 464-466.

Katritzky A. R. [et al.] Quantitative Measures of Solvent Polarity [Journal] // Chemical Reviews. - 2004. - 1 : Vol. 104. - pp. 175-198.

Kenny P. W., Montanari C. A. and Prokopczyk I. M. ClogPalk: a method for predicting alkane/water partition coefficient [Journal] // J Comput Aided Mol Des.. - 2013. - 1 : Vol. 5. - pp. 389-402.

Koba M., Bączek T. and Marszałł M. P. Importance of retention data from affinity and reverse-phase high-performance liquid chromatography on antitumor activity prediction of imidazoacridinones using QSAR strategy [Journal] // Journal of Pharmaceutical and Biomedical Analysis. - 2012. - Vols. 64-65. - pp. 87-93.

Koike M [et al.] Time-dependent elimination of cinoxacin in rats [Journal] // Journal of Pharmaceutical Sciences. - 1984. - 12 : Vol. 73. - pp. 1697-700.

Kowalski B. R. and Geladi P. Partial Least-Squares: A tutorial [Journal] // Analytica Chimica Acta. - [s.l.] : Elsevier Science Publishers, 1986. - Vol. 185. - pp. 1-17.

Kubinyi H. Comparative Molecular Field Analysis (CoMFA) [Journal] // The Encyclopaedia of Computational Chemistry. - 1998. - Vol. 1. - pp. 448-460.

Kubinyi H. Free Wilson Analysis. Theory, Application's and its Relationship to Hansch Analysis [Journal] // Quantitative Structure-Activity Relationships. - 1988. - Vol. 7. - pp. 121-133.

Kubinyi H. QSAR and 3D QSAR in drug design. Part 1: methodology [Journal] // Drug Discovery Today. - 1997. - 11 : Vol. 2.

Laane C. [et al.] Rules for Optimisation of Biocatalysis in Organic Solvents [Journal] // Biotechnology and Bioengineering. - 1987. - Vol. 30. - pp. 81-87.

Lamarche O., Platts J. A. and Hersey A. Theoretical prediction of partition coefficients via molecular electrostatic and electronic properties [Journal] // Journal of Chemical Information Computer Sciences. - 2004. - Vol. 44. - pp. 848–855 .

Leahy D. E. [et al.] Model solvent systems for QSAR Part 2 Fragment values (f-values) for the critical quartet [Journal] // Journal of the Chemical Society, Perkin Transactions 2. - 1992. - 2. - pp. 723-731.

Leo A. J. [et al.] Calculation of hydrophobic constant (log P) from π- and f- constants [Journal] // Journal Medicinal Chemistry. - 1975. - Vol. 18. - pp. 865-868.

Leo A., Hansch C. and Elkins D. Partition Coefficient and their uses [Journal]. - [s.l.] : Chemical reviews, 1971. - 6 : Vol. 71. - pp. 525-616.

Lesellier E. pider diagram: A universal and versatile approach for system comparison and classification Application to solvent properties [Journal] // Journal of Chromatography A. - 2015.

Levin V. A. Relationship of octanol/water partition coefficient and molecular weight to rat brain capillary permeability [Journal] // Journal of Medicinal Chemistry. - 1980. - 6 : Vol. 23. - pp. 682–684.

Li Z. [et al.] Property Calculation and Prediction for Selecting Solvent Systems in CCC [Journal] // Journal of Liquid Chromatography & Related Technologies. - 2003. - 9&10 : Vol. 26. - pp. 1397–1415.

Livingstone D. J. [et al.] Simultaneous Prediction of Aqueous Solubility and Octanol-Water Partition Coefficient for Pesticides Based on their Molecular Structure [Journal] // Journal of Computer-Aided Molecular Design. - 2001. - Vol. 15. - pp. 741–752.

Livingstone D. J. Theoretical Property Predictions [Journal] // Current Topics in Medicinal Chemistry. - 2003. - Vol. 3. - pp. 1171-1192.

Lombardo F. [et al.] ElogDoct: A Tool for Lipophilicity Determination in Drug Discovery. 2. Basic and Neutral Compounds [Journal] // Journal of Medicinal Chemistry. - 2001. - Vol. 44. - pp. 2490-2497.

Lu Y. [et al.] Screening of Complex Natural Extracts by Countercurrent Chromatography Using a Parallel Protocol [Journal] // Anal. Chem. . - 2009. - Vol. 81. - pp. 4048–4059.

Margraff R. and Foucault (Ed.) A. Centrifugal Partition Chromatography [Journal] // Chromatographic Science Series. - [s.l.] : Marcel Dekker: NewYork,, 1994. - Vol. 68. - pp. 331-350..

Martin A., Newburger J. and Adjei A. Extended Hildebrand Solubility Approach: Solubility of theophylline in polar binary solvents [Journal] // Journal of Pharmaceutical Sciences. - 1980. - Vol. 69. - pp. 487-491.

Menet J.-M. [et al.] Classification of Countercurrent Chromatography Solvent Systems on the Basis of the Capillary Wavelength [Journal] // Analytical Chemistry. - 1994. - Vol. 66. - pp. 168-176.

Meyer H. Zur Theorie der Alkoholnarkose [Journal] // Archiv für experimentelle Pathologie und Pharmakologie. - 1899. - 2-4 : Vol. 42. - pp. 109-118.

Michel T., Destandau E. and Elfakir C. New advances in Countercurrent chromatography and centrifugal partition chromatography: focus on coupling strategy [Journal] // Analytical and Bioanalytical Chemistry. - 2013. - Vol. 406. - pp. 957-969.

Mills I. [et al.] IUPAC Quantities, Units and Symbols in Physical Chemistry [Journal]. - [s.l.] : Blackwell Scientific Publications, 1993. - Vol. 2nd ed.

Murayama W. [et al.] A new centrifugal counter-current chromatograph and its application [Journal] // Journal of Chromatography. - 1982. - 1 : Vol. 239. - pp. 643-649..

Niewiadomy A. [et al.] Reversed Phase High-Performance Liquid Chromatography in Quantitative Structure-Activity Relationship Studies of New Fungicides [Journal] // Journal of Chromatography A. - 1998. - Vol. 828. - pp. 431-438.

Nys G. G. and Rekker R. F. Concept of hydrophobic fragmental constants (f-values) II. Extension of its applicability to the calculation of lipophilicities of aromatic and heteroaromatic structures [Journal] // Europe Journal of Medicinal Chemistry. - 1974. - Vol. 9. - pp. 361-375.

Oka F., Oka H. and Ito Y. Systematic search for suitable two-phase solvent systems for high speed counter-current chromatography [Journal] // Journal of Chromatography A. - 1991. - 1 : Vol. 538. - pp. 99-108.

Patil R. R. and Bari S. B. Pharmacophore-base 3D-QSAR study of fungal chitin synthesis inhibitors [Journal] // Medicinal Chemistry Research. - 2013. - Vol. 22. - pp. 1762-1772.

Petrauskas A. A. and Kolovanov, E. A. ACD/Log P method description [Journal] // Perspectives in Drug Discovery and Design. - 2000. - Vol. 19. - pp. 99-116.

Platts J. A. and Saunders R. A. Scaled polar surface area descriptors: development and application to three sets of partition coefficients [Journal] // New Journal of Chemistry. - 2004. - Vol. 28. - pp. 166–172 .

Poppe E. and Slaats H. Some observations on the derivations of solvent polarity factors from gas-liquid partition coefficients [Journal] // Chromatographia. - 1981. - Vol. 14. - pp. 89-94.

Raevsky O. A. [et al.] SLIPPER-2001 − Software for Predicting Molecular Properties on the Basis of Physicochemical Descriptors and Structural Similarity [Journal] // Journal of Chemical Information and Computer Sciences. - 2002. - 3 : Vol. 42. - pp. 540-549.

Rainsford K. D. Nimesulide - Actions and Uses [Journal] // Birkhäuser Basel. - 2005.

Reichardt C. Empirical Parameters of Solvent Polarity as Linear Free-Energy Relationships [Journal] // Angewandte Chemie International Edition in English. - 1979. - 2 : Vol. 18. - pp. 98-110.

Reichardt C. Empirical Parameters of the Polarity of Solvents [Journal] // Angewandte Chemie International Edition in English. - 1965. - 1 : Vol. 4. - pp. 29-40.

Reichardt C. Solvatochromic Dyes as Solvent Polarity Indicators [Journal] // Chemical Reviews. - 1994. - Vol. 94. - pp. 2319-2358.

Ren D.-B. [et al.] Correlation and prediction of partition coefficient using non-random two-liquid segment activity coefficient model for solvent system selection in counter-current chromatography separation [Journal] // Journal of Chromatography A. - 2013. - Vol. 1301. - pp. 10-18.

Rodgers S. L. [et al.] Predictivity of Simulated ADME AutoQSAR Models over Time [Journal] // Molecular Informatics. - 2011. - Vol. 30. - pp. 256 – 266.

Roy P. P. and Roy K. On Some Aspects of Variable Selection for Partial Least Squares Regression Models [Journal] // QSAR and Combinatorial Science. - 2008. - 3 : Vol. 27. - pp. 302-313.

Salum L. B. and Andricopulo A. D. Fragment-Based QSAR: Perspectives in Drug Design [Journal] // Molecular Diversity. - 2009. - Vol. 13. - pp. 277–285.

Schapire R. E. [et al.] Boosting the Margin: a New Explanation for the Effectiveness of Voting Methods [Journal] // The Annals of Statistics. - 1998. - 5 : Vol. 26. - pp. 1651-1686.

Schefzick S. [et al.] Comparative Molecular Field Analysis of Quinine Derivatives Used as Chiral Selectors in Liquid Chromatography: 3D QSAR for the Purposes of Molecular Design of Chiral Stationary Phase [Journal] // Chirality. - 2000. - Vol. 12. - pp. 742-750.

Selassie C. and Verma R. P. History of Quantitative Structure–Activity Relationships [Journal] // Burger's Medicinal Chemistry and Drug Discovery. - 2010. - pp. 1–96..

Skalicka-Woźniak K. and Garrard I. A comprehensive classification of solvent systems used for natural product purifications in countercurrent and centrifugal partition chromatography [Journal] // Natural product reports. - 2015.

Snyder R. L. Principles of adsorption chromatography: the separation of nonionic organic compounds [Journal] // Marcel Dekker. - 1968. - Vol. 3. - p. 167.

Sørensen J. M. and Arlt W. Liquid-liquid Equilibrium Data Collection: Ternary and Quaternary Systems [Journal] // Dechema. - 1980.

Sporna-Kucab A. Betalain Pigments Purification from Beetroot Juice [Journal] // Internal Report. - 2014.

Sumner N. Developing counter current chromatography to meet the needs of pharmaceutical discovery [Journal] // Journal of Chromatography A. - 2011. - Vol. 1218. - pp. 6107-6113.

Sutherland I. A. [et al.] Review of Progress Toward the Industrial Scale-Up of CCC [Journal] // Journal of Liquid Chromatography & Related Technologies. - 2005. - 12-13 : Vol. 28.

Sutherland I. A. and Fisher D. Role of Counter-Current Chromatography in the Modernisation of Chinese Herbal Medicine [Journal] // Journal of Chromatography A. - 2009. - Vol. 1216. - pp. 740-753.

Sutherland I. A. Liquid Stationary Phase Retention and Resolution in Hydrodynamic CCC [Journal] // Countercurrent Chromatography: The Support-Free Liquid Stationary Phase by A.Berthod. - 2002. - Vol. XXXVIII. - pp. 159-176.

Taft R. W. [et al.] Linear Solvation Energy Relationships. 29. Solution Properties of Some Tetraalkylammonium Halide Ion Pairs and Dissociated Ions [Journal] // Journal of the American Chemical Society. - 1985. - 11 : Vol. 107. - pp. 3105-3110.

Taft R. W. Polar and Steric Substituent Constants for Aliphatic and o-Benzoate Groups from Rates of Esterification and Hydrolysis of Esters [Journal] // Journal of the American Chemical Society. - 1952. - 12 : Vol. 74. - pp. 3120–3128.

Tonnesen H. H. Photostability of Drugs and Drug Formulations, Second Edition [Journal] // CRC Press. - 2004. - p. 30.

Umetrics www.umetrics.com/sites/default/files/kb/multivariate_faq.pdf [Journal] // Additional SIMCA documentation. - 2015.

Vapnik V. and Lerner A. Pattern Recognition Using Generalised Portrait Method [Journal] // Automation and Remote Control. - 1963. - Vol. 24. - pp. 774-780.

Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules [Journal] // Journal of Chemical Information and Modelling. - 1988. - 1 : Vol. 28. - pp. 31-36.

Wong W. W. L. and Burkowski F. J. A constructive approach for discovering new drug leads: Using a kernel methodology for the inverse-QSAR problem [Journal] // Journal of Cheminformatics. - 2009. - 4 : Vol. 1.

Wood D. J. [et al.] Automated QSAR with a Hierarchy of Global and Local Models [Journal] // Mol. Inf.. - 2011. - Vol. 30. - pp. 960-972.

Wood P. L. The Hydrodynamics of Countercurrent Chromatography in J-Type Centrifuges [Journal] // PhD Thesis, Brunel University, UK. - 2002.

Yan C. [et al.] QSAR correlation of the melting points for imidazolium bromides and imidazolium chlorides ionic liquids [Journal] // Fluid Phase Equilibria. - 2010. - Vol. 292. - pp. 104-109.

Zaslavasky B. Y. Aqueous two-phase partitioning: Physical chemistry and bioanalytical applications [Journal] // Marcel Dekker Inc, New York.. - 1995.

Zerara M. and Brickmann J. Parameterisation of an empirical model for the prediction of n-octanol, alkane and cyclohexane/water as well as brain/blood partition coefficients [Journal] // Journal of Computer-Aided Molecular Design. - 2009. - Vol. 23. - pp. 105-111.

Zissimos A. M [et al.] A comparison between the Two General Sets of Linear Free Energy Descriptors of Abraham and Klamt [Journal] // Journal of Chemical Information and Modelling. - 6 : Vol. 42. - pp. 1320-1331.

9. Appendix

9.1. Experimentally determined logK_d values used to train the initial QSAR models

Compound Name	Experimentally determined logK_d values in 6 different HEMWat systems					
	8	14	17	20	22	26
3-Bromobenzoic acid	2.40	1.20	0.19	-0.61	-0.62	-0.88
3-Hydroxybenzoic acid	1.03	-0.15	-1.35	-1.55	NR	-2.61
Aspirin	1.30	0.21	-0.73	-1.28	NR	NR
Cinoxacin	-2.68	-2.88	-3.52	NR	NR	NR
Glyburide	2.20	0.90	-0.75	-2.02	-2.36	-2.51
Hesperidin	-1.43	-2.74	NR	NR	NR	NR
Hydrocortisone-21-hemisuccinate	1.64	-0.09	-0.62	NR	NR	NR
Ibuprofen	NR	1.83	0.62	-0.04	-0.28	-0.61
Naproxen	2.65	1.24	0.00	-0.65	-0.86	NR
Sulfacetamide	0.11	-0.94	-2.14	-3.26	-3.62	-3.74
Sulfamethoxazole	1.34	-0.44	-2.22	-3.95	NR	NR
Sulfathiazole	-0.21	NR	NR	NR	NR	NR
Sulfisoxazole	-0.12	NR	NR	NR	NR	NR
Tolbutamide	2.24	0.79	-0.44	-1.30	-1.71	-1.84
Haloperidol	3.06	1.61	0.25	-0.70	-0.89	NR
Lidocaine	1.92	1.13	0.22	-0.40	-0.62	-0.86
Quinine	1.51	0.48	-0.53	-0.97	-1.11	-1.98
Reserpine	2.46	1.49	-0.10	NR	NR	NR
Trimethoprim	-0.05	-1.12	NR	NR	NR	NR
Tryptamine	0.28	-0.26	-1.03	-1.93	-2.24	-2.26
1,2-Dihydroxynaphthalene	2.16	0.51	-1.14	-2.42	-2.81	-2.30
1,2-dimethyl-1H-indole-3-carboxaldehyde	1.54	0.14	-0.44	-1.78	-1.83	NR
1,3-diacetylindole	2.14	1.05	0.09	-0.65	-0.87	-1.17
2-ethylanthraquinone	3.46	2.20	1.22	0.74	0.55	0.19
4-bromochlorobenzene (p-Chlorobromobenzene)	2.56	2.13	1.38	1.05	0.80	0.50
4-hydroxyphenylacetamide	-0.48	-1.24	-2.30	-3.20	-3.27	NR
4-nitroaniline	1.91	0.69	-0.43	-1.30	-1.77	NR

Albendazole	2.58	1.18	-0.02	-0.73	-0.97	NR
Aminopyrine	0.08	-0.66	-1.27	-1.92	NR	NR
Antipyrine	-0.46	-1.11	-1.72	-2.19	NR	NR
Benzoquinone	-0.67	-0.94	-1.37	-1.91	-1.92	-2.04
Biphenyl	2.41	1.86	1.26	0.87	0.71	0.38
Caffeine	-0.34	-0.84	-1.42	-1.86	-1.96	-1.92
Clofazimine	3.16	2.34	1.89	1.48	1.04	0.44
Cortexolone (Reichstein's substance)	0.52	-1.18	-2.22	NR	NR	NR
Cyclododecanone	NR	NR	NR	0.79	0.59	NR
Diethylstilbestrol	2.35	1.35	-0.02	-1.09	-1.49	-2.10
Diphenyl sulfone	2.45	1.08	0.03	-0.49	-0.62	-0.71
Dipyridamole	1.43	-0.25	-1.59	-2.70	-3.00	-4.35
Griseofulvin	1.85	0.34	-0.91	-1.74	-1.99	-2.13
indole-5-methanol	1.00	-0.31	-1.03	-1.89	-2.14	-2.45
methyl 2-acetamidobenzoate	1.64	0.86	0.09	-0.39	-0.51	-0.68
Methyl-2-Amino-5-Bromobenzoate	2.53	1.18	0.50	-0.04	-0.21	-0.47
Methyl-4-amino-3-methylbenzoate	1.87	0.33	-1.42	-1.59	-1.81	-1.90
Napthalene	2.61	1.91	1.15	0.79	0.67	-1.46
Pentoxifylline	1.53	0.33	-1.34	NR	NR	NR
Phenanthrene	NR	2.09	1.36	0.96	0.80	0.42
Phenothiazine	3.67	1.96	0.91	0.34	0.04	-0.51
Phenylbutazone	2.43	1.67	0.76	-0.07	-0.32	-0.63
Pyrene	3.79	2.26	1.49	1.10	0.89	0.52
Theophylline	-0.53	-0.92	NR	-1.29	-2.32	-2.07
Thiamphenicol	-0.75	NR	NR	NR	NR	NR
Warfarin	2.80	0.93	-0.75	-1.49	-1.69	-1.95

9.2. Top14 AZ descriptors

ACDlogD65	ACDlogD65 is calculated as the octanol/water distribution coefficient at pH 6.5
ACDlogD74	ACDlogD74 is calculated as the octanol/water distribution coefficient at pH 7.4.
ACDlogP	ACDlogP is calculated as the octanol/water partition coefficient for the neutral species.
ClogP	ClogP is a predicted octanol/water partition coefficient from Daylight/Biobyte
HBA	Lipinski number of HB acceptors = number of O+N.
HBD	Lipinski number of HB donors = number of OH+NH.
PSA	Van der Waals radius surface, summed over all N, O and attached hydrogens, 1-3 overlap correction.
RotBond	Number of non-terminal flexible bonds.
VOL	Gaussian volume. A measure of molecular volume.
MW	Molecular weight calculated by the OEChem toolkit
NPSA	Total surface area minus polar surface area: AREA-PSA
IonClass	Acid, Base, Neutral or Zwitterion
RingCount	Number of rings (smallest set of smallest rings).
Lipinski	Number of failures at Lipinski's rule of 5.

9.3. 196 AZ descriptors

Lipophilicity	
ACDlogD6.5	ACDlogD65 is calculated as the octanol/water distribution coefficient at pH 6.5
ACDlogD7.4	ACDlogD74 is calculated as the octanol/water distribution coefficient at pH 7.4.
ACDlogP	ACDlogP is calculated as the octanol/water partition coefficient for the neutral species.
GClogP	Octanol/water partition coefficient based on Ghose/Crippen atom types
NNlogP	Octanol/water partition coefficient using a neural network approach based on Ghose/Crippen atom types
ClogP	ClogP is a predicted octanol/water partition coefficient from Daylight/Biobyte
H-Bonding	
HBA_nonLipinski	Number of potential H-bond acceptor bonds, not Lipinski definitions.
HBD_nonLipinski	Number of potential H-bond donor bonds, not Lipinski definitions.
HBA	Lipinski number of HB acceptors = number of O+N.
HBD	Lipinski number of HB donors = number of OH+NH.
HB_sum	Total number of potential H-bonds using Lipinski definition: HBA+HBD
HBA_Selma	Number of hydrogen bond acceptors.
HBD_Selma	Number of hydrogen bond donors.
HBA_Raevsky	Number of hydrogen bond acceptors according to Raevsky (HYBOT).
HBD_Raevsky	Number of hydrogen bond donors according to Raevsky (HYBOT).

HBAmax	Highest free energy factor for H-bond acceptors according to Raevsky (HYBOT).
HBDmax	Highest free energy factor for H-bond donors according to Raevsky (HYBOT).
HBAsum	Sum of acceptor free energies according to Raevsky (HYBOT).
HBDsum	Sum of donor free energies according to Raevsky (HYBOT).
HBsumTotal	Sum of donor and acceptor free energies according to Raevsky (HYBOT).
PSA	Van der Waals radius surface, summed over all N, O and attached hydrogens, 1-3 overlap correction.
SAS_HB_A_AREA	Solvent accessible surface H-bond acceptor area.
SAS_HB_D_AREA	Solvent accessible surface H-bond donor area.
SAS_POL_AREA	Solvent accessible surface polar area.
SPEC_HB_TOT	HBsum/HeavyAtomCount.
SPEC_SAS_HB_A_AREA	SAS_HB_A_AREA / SAS_TOT_AREA.
SPEC_SAS_HB_D_AREA	SAS_HB_D_AREA / SAS_TOT_AREA.
SPEC_SAS_POL_AREA	SAS_POL_AREA / SAS_TOT_AREA.
SPEC_VDW_HB_A_AREA	VDW_HB_A_AREA / VDW_AREA.
SPEC_VDW_HB_D_AREA	VDW_HB_D_AREA / VDW_AREA.
SPEC_VDW_POL_AREA	VDW_POL_AREA / VDW_AREA
VDW_HB_A_AREA	Van der Waals H-bond acceptor area.
VDW_HB_D_AREA	Van der Waals H-bond donor area.
VDW_POL_AREA	Van der Waals polar surface area.
PSA_percentage	Percent polar surface area: PSA/AREA
Size / Shape	
CMR	Calculated molar refractivity. Largely a volume descriptor, highly correlated with molecular weight.
RotBond	Number of non-terminal flexible bonds.

VOL	Gaussian volume. A measure of molecular volume.
GraphDiameter	Longest of the shortest topological paths between center of the molecule and other atoms.
GraphRadius	Longest of the shortest topological paths between atoms.
M1M	Moment of inertia along the first principal axis of the molecule.
M2M	Moment of inertia along the second principal axis of the molecule.
M3M	Moment of inertia along the third principal axis of the molecule.
MolVol2D	Van der Waals radius based volume, summed over all atoms with a 1-3 overlap correction.
MW	Molecular weight calculated by the OEChem toolkit
MolFlex	Molecular flexibility, calculated as 2.85 ^ (sp3-count + 0.5*sp2 count + 0.5* nb_rings count -1)
VDW_AREA	Van der Waals molecular surface area.
NPSA	Total surface area minus polar surface area: AREA-PSA
AREA	Van der Waals radius surface, summed over all atoms, with a 1-3 overlap correction.
NPSA_percentage	Percent non-polar surface area: NPSA/AREA
OVAL_NEW	TSA / the area of a sphere with the volume given by MolVol2D
SAS_NONPOL_AREA	Solvent accessible surface non-polar area.
SAS_TOT_AREA	Solvent accessible surface total area.
SPEC_FLEX_BND	Defined as ratio FLEX_BND/HEAVIES.
SPEC_SAS_NONPOL_AREA	SAS_NONPOL_AREA / SAS_TOT_AREA.
SPEC_VDW_NONPOL_AREA	VDW_NONPOL_AREA / VDW_AREA.
VDW_NONPOL_AREA	Van der Waals non-polar surface area.
Charge / Polarity	

FractionNeutral	10 ^ (ACDlogD74 - ACDlogP)
FractionIonized	(1 - FractionNeutral)
Acid	Presence of an acid function.
Base	Presence of a basic function.
Neutral	Absence of neither basic nor acid functions.
Zwitterion	Presence of at least one acid and basic function.
IonClass	Acid, Base, Neutral or Zwitterion
AverNegCharge_GM	Average negative charge using the Gasteiger-Marsili partial charge equilibration.
AverNegCharge_GH	Average negative charge using the Gasteiger-Huckel partial charge equilibration.
AverPosCharge_GM	Average positive charge using the Gasteiger-Marsili partial charge equilibration.
AverPosCharge_GH	Average positive charge using the Gasteiger-Huckel partial charge equilibration.
ChargeRang_GM	Maximum minus minimum charge using Gasteiger-Marsili partial charge equilibration.
ChargeRange_GH	Maximum minus minimum charge using Gasteiger-Huckel partial charge equilibration.
CHARGED	= 1 if CHARGES > 0.
CHARGES	POS_charges + NEG_charges.
DipoleMomGH	Largest of all products of inter-atomic topological distances multiplied with the Gasteiger-Huckel charge range.
DipoleMomGM	Largest of all products of inter-atomic topological distances multiplied with the Gasteiger-Marsili partial charges
HOMO	Huckel molecular orbitals, Highest occupied molecular orbital energy.
LUMO	Huckel molecular orbitals, Lowest unoccupied molecular orbital energy.
HuckelPiEnergy	Huckel molecular orbitals, pi electrons energy.
HuckelResEnergy	Huckel molecular orbitals, resonance energy.

MaxNegChargeGM	Maximum negative charge using the Gasteiger-Marsili partial charge equilibration.
MaxPosChargeGM	Maximum positive charge using the Gasteiger-Marsili partial charge equilibration.
MaxNegChargeGH	Maximum negative charge using the Gasteiger-Huckel partial charge equilibration.
MaxPosChargeGH	Maximum positive charge using the Gasteiger-Huckel partial charge equilibration.
MM_FHADSA	A measure of the dispersion of the charge on hydrogen bond donor and acceptor atoms on the surface.
MM_FHASA	A measure of the dispersion of the charge on hydrogen bond acceptor atoms on the surface.
MM_FHDSA	A measure of the dispersion of the charge on hydrogen bond donor atoms on the surface.
MM_HACA	A measure of the dispersion of the charge on hydrogen bond acceptor atoms on the surface.
MM_HADCA	A measure of the dispersion of the charge on hydrogen bond donor and acceptor atoms on the surface.
MM_HADSA	A measure of the dispersion of the charge on hydrogen bond donor and acceptor atoms on the surface.
MM_HASA	A measure of the dispersion of the charge on hydrogen bond acceptor atoms on the surface.
MM_HDCA	A measure of the dispersion of the charge on hydrogen bond donor atoms on the surface.
MM_HDSA	A measure of the dispersion of the charge on hydrogen bond donor atoms on the surface.
MM_MAXNEG	Maximum negative atomic charge.
MM_MAXPOS	Maximum positive atomic charge.
MM_PCWT	Most negative partial charge weighted topological index.
MM_QMIN	Charge of the most negative atom

MM_QC	Sum of atomic charges on C.
MM_QH	Sum of atomic charges on H.
MM_QN	Sum of atomic charges on N.
MM_QnegMean	Mean of negative charges.
MM_QnegVar	Variance of negative charges.
MM_QO	Sum of atomic charges on O.
MM_QON	Sum of atomic charges on O+N.
MM_QposMean	Mean of positive charges.
MM_QposVar	Variance of positive charges.
MM_RNCS	Relative negative charge surface area.
MM_SAS_EP_N_AREA	Area of solvent accessible surface with negative electrostatic potential.
MM_SAS_EP_N_MEAN	Mean of negative electrostatic potentials on solvent accessible surface.
MM_SAS_EP_N_SUM	Sum of negative electrostatic potentials on solvent accessible surface.
MM_SAS_EP_N_VAR	Variance of negative electrostatic potentials on solvent accessible surface.
MM_SAS_EP_P_AREA	Area of solvent accessible surface with positive electrostatic potential.
MM_SAS_EP_P_MEAN	Mean of positive electrostatic potentials on solvent accessible surface.
MM_SAS_EP_P_SUM	Sum of positive electrostatic potentials on solvent accessible surface.
MM_SAS_EP_P_VAR	Variance of positive electrostatic potentials on solvent accessible surface.
MM_SPEC_SAS_EP_N_AREA	Proportion of negative electrostatic potential on the solvent accessible surface area.
MM_SPEC_SAS_EP_P_AREA	Proportion of positive electrostatic potential on the solvent accessible surface area.
MM_SPEC_VDW_EP_N_AREA	Proportion of negative electrostatic potential on the Van der Waals surface area.
MM_SPEC_VDW_EP_P_AREA	Proportion of positive electrostatic potential on the Van der Waals surface area.

MM_VDW_EP_N_AREA	Area of Van der Waals surface with negative electrostatic potential.
MM_VDW_EP_N_MEAN	Mean of negative electrostatic potentials on Van der Waals surface.
MM_VDW_EP_N_SUM	Sum of negative electrostatic potentials on Van der Waals surface.
MM_VDW_EP_N_VAR	Variance of negative electrostatic potentials on Van der Waals surface.
MM_VDW_EP_P_AREA	Area of Van der Waals surface with positive electrostatic potential.
MM_VDW_EP_P_MEAN	Mean of positive electrostatic potentials on Van der Waals surface.
MM_VDW_EP_P_SUM	Sum of positive electrostatic potentials on Van der Waals surface.
MM_VDW_EP_P_VAR	Variance of positive electrostatic potentials on Van der Waals surface.
NegIonCenters	Number of negative ionisation centres.
NEG_charges	Number of acidic groups likely to be ionised at pH 7.4.
NEL_CHNOS	Number of electrons for C, H, N, O, S atoms.
NEL_all	Number of electrons including all atoms.
Polarizability	Bobby Glenn's polarizability scheme (modified from Pauling).
PosIonCenters	Number of positive ionisation centres.
POS_charges	Number of basic groups likely to be ionised at pH 7.4.
Atom Counts	
Amine1	Number of primary amines.
Amine2	Number of secondary amines.
Amine3	Number of tertiary amines.
AromCount	Number of aromatic atoms.
AtomCount	Total number of atoms, including hydrogens.
BromineCount	Number of bromine atoms.

CarbonCount	Number of carbons.
ChlorineCount	Number of chlorine atoms.
FluorineCount	Number of fluorine atoms.
HAROM	Number of hydrogens linked to an aromatic atom.
HeavyAtomCount	Number of non-hydrogen atoms.
IodineCount	Number of iodine atoms.
MaxRing1	Size of the largest ring cycle.
MaxRing2	Size of 2nd largest ring cycle.
MaxRing3	Size of 3nd largest ring cycle.
MWNPat	MW * NPat/AT_TOT Proportion of MW accounted for by the excess of non-polar atoms (by number)
MWPat	MW *Pat/AT_TOT Proportion of MW accounted for by the polar atoms (by number).
MWSHDA	MW * HB_TOT/AT_TOT Proportion of MW accounted for by hydrogen bonding groups (by number)
OxygenCount	Number of oxygens.
NitrogenCount	Number of nitrogens.
NHCount	Number of NH groups
OHCount	Number of OH groups
PhosphorusCount	Number of phosphorous atoms.
SulfurCount	Number of sulphur atoms.
HalogenCount	Number of halogen atoms.
SiliconCount	Number of silicon atoms.
RingCount	Number of rings (smallest set of smallest rings).
NonpolarCount	Sum of (carbons, halogens) minus polar count.
NonpolarCountMW	Nonpolar count divided by molecular weight
NPAT	Number of excess of non-polar atoms (NBC + NBX - PAT).
BondCount	Number of bonds (all bonds).

RigidbondCount	Number of bonds - number of rotatable bonds - number of terminal bonds.
PAT	Number of polar atoms (O, N, S, P).
PolarCount	Sum of N, P, O, S atoms where P, S are in high oxidation state.
PolarCountMW	Polar count divided by molecular weight
PIAT	Number of pi atoms (number of atoms linked to double bonds + number of halogen atoms).
QUATER	Number of quaternary nitrogen.
Topology	
Balaban	Topological distance matrix based index related to ring structures.
IC	Shannon entropy (zero for symmetrical molecules; higher in non-symmetrical molecules).
Kappa1	Topological index.
Kappa2	Topological index.
Kappa3	Topological index.
Chi0	Sum of reciprocal square roots of valences over all atoms.
Chi2	Sum of reciprocal square roots of valences over all triplet pairs.
Chi3c	Sum of reciprocal square roots of valences over all 4-count branched atom paths.
Chi3p	Sum of reciprocal square roots of valences over all 4-count linear atom paths.
Chi4c	Sum of reciprocal square roots of valences over all 5-count branched atom paths.
Chi4p	Sum of reciprocal square roots of valences over all 5-count linear atom paths.
Chi5c	Sum of reciprocal square roots of valences over all 6-count branched atom paths.
Chi5p	Sum of reciprocal square roots of valences over all 6-count linear atom paths.

Chi6p	Sum of reciprocal square roots of valences over all 7-count linear atom paths.
MaxEV1	Largest maximum eigenvalue from connectivity matrix, where diagonal has atomic weights.
MaxEV2	2nd largest maximum eigenvalue from connectivity matrix, where diagonal has atomic weights.
MaxEV3	3rd largest maximum eigenvalue from connectivity matrix, where diagonal has atomic weights.
MinEV1	Smallest minimum eigenvalue from connectivity matrix, where diagonal has atomic weights.
MinEV2	2nd smallest minimum eigenvalue from connectivity matrix, where diagonal has atomic weights.
MinEV3	3rd smallest minimum eigenvalue from connectivity matrix, where diagonal has atomic weights.
MindistAA	Shortest H-bond acceptor-acceptor distance (not in the same moiety).
MindistDA	Shortest H-bond donor-acceptor distance (not in the same moiety).
MindistDD	Shortest H-bond donor-donor distance (not in the same moiety).
Motoc	Topological distance matrix based index related to ring structures.
Randic	Topological distance matrix based index related to ring structures.
Wiener	Half sum of the topological distance matrix of connectivity.
SIC	Structural information content of 0 order.
Druggability	
Lipinski	Number of failures at Lipinski's rule of 5.

9.4. Descriptors that can be manually calculated or obtained from freeware

ACDlogP	ACDlogP is calculated as the octanol/water partition coefficient for the neutral species.
HBA	Lipinski number of HB acceptors = number of O+N.
HBD	Lipinski number of HB donors = number of OH+NH.
HB_sum	Total number of potential H-bonds using Lipinski definition: HBA+HBD
HBA_Selma	Number of hydrogen bond acceptors.
HBD_Selma	Number of hydrogen bond donors.
SPEC_HB_TOT	HBsum/HeavyAtomCount.
CMR	Calculated molar refractivity. Largely a volume descriptor, highly correlated with molecular weight.
RotBond	Number of non-terminal flexible bonds.
VOL	Gaussian volume. A measure of molecular volume.
MW	Molecular weight calculated by the OEChem toolkit
Amine1	Number of primary amines.
Amine2	Number of secondary amines.
Amine3	Number of tertiary amines.
AromCount	Number of aromatic atoms.
AtomCount	Total number of atoms, including hydrogens.
BromineCount	Number of bromine atoms.
CarbonCount	Number of carbons.
ChlorineCount	Number of chlorine atoms.
FluorineCount	Number of fluorine atoms.
HAROM	Number of hydrogens linked to an aromatic atom.
HeavyAtomCount	Number of non-hydrogen atoms.
IodineCount	Number of iodine atoms.
MaxRing1	Size of the largest ring cycle.
MaxRing2	Size of 2nd largest ring cycle.
MaxRing3	Size of 3nd largest ring cycle.
OxygenCount	Number of oxygens.
NitrogenCount	Number of nitrogens.
NHCount	Number of NH groups

OHCount	Number of OH groups
PhosphorusCount	Number of phosphorous atoms.
SulfurCount	Number of sulphur atoms.
HalogenCount	Number of halogen atoms.
SiliconCount	Number of silicon atoms.
RingCount	Number of rings (smallest set of smallest rings).
NonpolarCount	Sum of (carbons, halogens) minus polar count.
NonpolarCountMW	Nonpolar count divided by molecular weight
NPAT	Number of excess of non-polar atoms (NBC + NBX - PAT).
BondCount	Number of bonds (all bonds).
RigidbondCount	Number of bonds - number of rotatable bonds - number of terminal bonds.
PAT	Number of polar atoms (O, N, S, P).
PolarCount	Sum of N, P, O, S atoms where P, S are in high oxidation state.
PolarCountMW	Polar count divided by molecular weight
PIAT	Number of pi atoms (number of atoms linked to double bonds + number of halogen atoms).
QUATER	Number of quaternary nitrogen.
Lipinski	Number of failures at Lipinski's rule of 5.
A	Hydrogen Bonding Acidity
B	Hydrogen Bonding basicity
S	Polarisability
E	Molar Refractivity
V	McGowan Volume

9.5. The 20 descriptors with the largest coefficient values from the 196 AZ descriptors identified using PLS

9.5.1. HEMWat 8

ACDlogP
ClogP
NNlogP
ACDlogD65
GClogP
MMVDWEPNVAR
MMVDWEPNMEAN
HBAmax
ACDlogD74
MWNPat
AverNegChargeGM
NPAT
HAROM
PolarCountMW
MMQnegVar
NPSApercentage
PSApercentage
SPECHBTOT
MMHADCA
MMHASA

9.5.2. HEMWat 14

ClogP
ACDlogP
MMVDWEPNMEAN
GClogP
NNlogP
MMVDWEPNVAR
MMSASEPNVAR
MaxNegChargeGH
HBAmax
MMSASEPPVAR
ACDlogD65
HAROM
MMSASEPNMEAN
MMSASEPPMEAN
AverNegChargeGM
NitrogenCount
PIAT
PolarCountMW
MWNPat
ACDlogD74

9.5.3. HEMWat 17

ClogP
ACDlogP
MMVDWEPNMEAN
MMSASEPNVAR
GClogP
NNlogP
MMSASEPNMEAN
MMVDWEPNVAR
MMSASEPPVAR
MMSASEPPMEAN
MWNPat
ACDlogD65
PIAT
HBAmax
HAROM
ACDlogD74
HalogenCount
MaxEV2
NPAT
AverNegChargeGM

9.5.4. HEMWat 20

MMVDWEPNMEAN
ClogP
MMSASEPNMEAN
ACDlogP
GClogP
NNlogP
ACDlogD65
MMSASEPNSUM
MWNPat
ACDlogD74
MaxRing1
MaxNegChargeGM
NPAT
AverNegChargeGM
NonpolarCount
HAROM
MWSHDA
MMMAXNEG
MMQMIN
HBDsum

9.5.5. HEMWat 22

MMVDWEPNMEAN
MMSASEPNMEAN
ClogP
ACDlogP
ACDlogD65
NNlogP
GClogP
PolarCountMW
SPECSASNONPOLAREA
SPECVDWNONPOLAREA
MMFHASA
MMFHADSA
PSApercentage
NPSApercentage
MMQnegVar
SPECSASHBAAREA
MMHASA
MMHADSA
SPECVDWHBAAREA
SPECHBTOT

9.5.6. HEMWat 26

MMVDWEPNMEAN
ClogP
ACDlogP
MindistDA
MWNPat
MaxNegChargeGH
ChargeRangeGH
DipoleMomGM
VDWHBAAREA
OHCount
SASHBAAREA
HBsumTotal
SASHBDAREA
SASPOLAREA
VDWPOLAREA
MMQnegVar
MMHACA
VDWHBDAREA
HBDSelma
SPECVDWHBAAREA

9.6. The 20 descriptors with the largest coefficient values from the 196 AZ descriptors and the five Abraham parameters identified using PLS

9.6.1. HEMWat 8

ACDlogP
ClogP
ACDlogD65
NNlogP
B
ACDlogD74
AverNegChargeGM
GClogP
PolarCountMW
HBsumTotal
MMHACA
AverNegChargeGH
NonpolarCountMW
MWSHDA
MMHASA
SPECHBTOT
NPSApercentage
PSApercentage
MMFHASA
MMFHADSA

9.6.2. HEMWat14

MMVDWEPNMEAN
ClogP
ACDlogP
GClogP
NNlogP
MMVDWEPNVAR
MMSASEPPMEAN
MMSASEPNVAR
MMSASEPPVAR
ACDlogD65
AverNegChargeGM
MMSASEPNMEAN
PolarCountMW
B
HAROM
MWNPat
MinEV1
HBAsum
ACDlogD74
PIAT

9.6.3. HEMWat 17

ClogP
ACDlogP
MMVDWEPNMEAN
MMSASEPNVAR
NNlogP
GClogP
MMSASEPNMEAN
MMVDWEPNVAR
MMSASEPPMEAN
MMSASEPPVAR
ACDlogD65
MWNPat
HBAmax
ACDlogD74
PIAT
HAROM
HalogenCount
MaxEV2
NPAT
AverNegChargeGM

9.6.4. HEMWat 20

MMVDWEPNMEAN
ClogP
MMSASEPNVAR
MMSASEPNMEAN
ACDlogP
MMSASEPPMEAN
MMVDWEPNVAR
GClogP
NNlogP
HBAmax
ACDlogD65
MMSASEPNSUM
MMSASEPPVAR
ACDlogD74
MWNPat
MMVDWEPPMEAN
MaxRing1
MaxNegChargeGM
SPECSASHBDAREA
NPAT

9.6.5. HEMWat 22

MMVDWEPNMEAN
MMSASEPPVAR
MMSASEPPMEAN
MMSASEPNMEAN
ClogP
ACDlogP
MMVDWEPPMEAN
MMVDWEPPVAR
ACDlogD65
SPECSASHBDAREA
SASHBDAREA
NNlogP
VDWHBDAREA
SPECVDWHBDAREA
GClogP
SASPOLAREA
HBDSelma
SASHBAAREA
VDWPOLAREA
MMHDCA

9.6.6. HEMWat 26

MMVDWEPNMEAN
ClogP
MaxNegChargeGH
ChargeRangeGH
SASHBAAREA
MaxPosChargeGH
VDWHBAAREA
HBsumTotal
DipoleMomGM
SASPOLAREA
SPECSASHBAAREA
OHCount
PAT
SPECVDWHBAAREA
VDWPOLAREA
MMHACA
SASHBDAREA
NonpolarCountMW
MMHASA
SPECSASHBDAREA

9.7. The coefficients and the corresponding descriptors for the QSAR equation

9.7.1. HEMWat 8

Descriptor	Coefficient
ACDlogD6.5	0.066
ACDlogD7.4	-0.042
ACDlogP	0.287
ClogP	0.308
Constant	0.265

9.7.2. HEMWat 14

Descriptor	Coefficient
ACDlogD6.5	0.035
ACDlogD7.4	0.023
ACDlogP	0.061
AromCount	6.39E-03
AverNegCharge_GH	-0.087
AverNegCharge_GM	1.032
AverPosCharge_GH	-0.206
AverPosCharge_GM	0.624
ChargeRang_GM	-0.042
ClogP	0.063
Constant	1.747
GClogP	0.056
HAROM	0.018
HBA	-4.90E-03
HBA_Raevsky	-6.49E-03
HBA_SELMA	-1.56E-02
HBAmax	-0.105
HBAsum	-0.015
HBD	-0.016
HBD_nonlipinski	-0.016
HBD_Raevsky	-0.016
HBD_SELMA	-0.031
HBDsum	3.39E-03
HBsum	2.54E-03
HBsumTotal	-9.10E-03
MAX_NEGCharge_GH	-0.242
MAX_NEGCharge_GM	0.226
MAX_POSCharge_GM	0.072
MM_FHADSA	7.50E-03
MM_FHASA	-4.28E-03
MM_FHDSA	1.71E-02
MM_HACA	-0.064
MM_HADCA	5.00E-06
MM_HADSA	0.277
MM_HASA	-0.849
MM_HDCA	-0.12
MM_HDSA	-0.045
MM_MAX_NEG	0.046
MM_MAX_POS	0.098
MM_QMIN	0.046
MM_Qneg_MEAN	-0.599
MM_Qneg_VAR	-0.197
MM_QON	-1.78E-03
MM_Qpos_VAR	0.164

MM_SAS_EP_N_MEAN	0.011
MM_SAS_EP_N_VAR	-0.018
MM_SAS_EP_P_MEAN	-0.011
MM_SAS_EP_P_VAR	-0.015
MM_VDW_EP_N_MEAN	7.83E-03
MM_VDW_EP_N_VAR	-9.90E-03
MWNPat	1.24E-03
MWPat	3.67E-04
MWSHDA	1.10E-03
NitrogenCount	-0.032
NNlogP	0.056
NonpolarCount	2.27E-03
NonpolarCountMW	-1.771
NPAT	5.86E-03
NPSA_percentage	-1.48E-03
PAT	-2.84E-03
PIAT	1.31E-02
PolarCountMW	7.655
PSA	2.90E-04
PSA_percentage	1.48E-03
SAS_HB_A_AREA	-2.71E-04
SAS_HB_D_AREA	-5.88E-04
SAS_NONPOL_AREA	2.75E-04
SAS_POL_AREA	-1.10E-04
SPEC_HB_TOT	0.117
SPEC_SAS_HB_A_AREA	-0.109
SPEC_SAS_HB_D_AREA	-0.2
SPEC_SAS_NONPOL_AREA	0.052
SPEC_SAS_POL_AREA	-0.052
SPEC_VDW_HB_A_AREA	-0.07
SPEC_VDW_HB_D_AREA	-0.142
SPEC_VDW_NONPOL_AREA	0.027
SPEC_VDW_POL_AREA	-0.027
VDW_HB_A_AREA	-2.97E-04
VDW_HB_D_AREA	-1.05E-03
VDW_POL_AREA	-9.13E-06

9.7.3. HEMWat 17

Descriptors	Coefficients
ACDlogD6.5	0.025
ACDlogD7.4	0.019
ACDlogP	0.045
AromCount	6.19E-03
AverNegCharge_GH	0.022
AverNegCharge_GM	0.667
AverPosCharge_GH	-0.205
AverPosCharge_GM	-0.396
ChargeRang_GM	-0.137
ChargeRange_GH	0.016
ClogP	0.046
Constant	0.832
DipoleMom_GM	-0.004
GClogP	0.041
HalogenCount	0.065
HAROM	0.012
HBA	-7.17E-04
HBA_Raevsky	-1.40E-03
HBA_SELMA	-6.07E-03
HBAmax	-0.064
HBAsum	-5.02E-03
HBD	-4.61E-03
HBD_nonlipinski	-4.61E-03
HBD_Raevsky	-4.61E-03
HBD_SELMA	-0.014
HBDsum	1.64E-03
HBsum	2.43E-03
HBsumTotal	-3.18E-03
MAX_NEGCharge_GH	-5.99E-02
MAX_NEGCharge_GM	0.263
MAX_POSCharge_GH	-8.28E-03
MAX_POSCharge_GM	-0.169
MaxEV2	0.021
MinEV1	-0.076
MM_FHADSA	-6.40E-03
MM_FHASA	-0.013
MM_FHDSA	-0.044
MM_HACA	-0.034
MM_HADCA	-0.012
MM_HADSA	-0.416
MM_HASA	-0.903
MM_HDCA	-0.188
MM_HDSA	-3.193
MM_MAX_NEG	0.04

MM_MAX_POS	1.39E-03
MM_QMIN	0.04
MM_Qneg_VAR	-0.169
MM_QO	5.99E-03
MM_QON	2.24E-03
MM_Qpos_VAR	-0.079
MM_SAS_EP_N_MEAN	0.011
MM_SAS_EP_N_VAR	-0.016
MM_SAS_EP_NSUM	0.016
MM_SAS_EP_P_MEAN	-9.52E-03
MM_SAS_EP_P_VAR	-0.012
MM_VDW_EP_N_MEAN	5.13E-03
MM_VDW_EP_N_VAR	-6.56E-03
MM_VDW_EP_P_VAR	-1.08E-03
MWNPat	1.16E-03
MWPat	4.64E-04
MWSHDA	7.38E-04
NNlogP	0.041
NonpolarCount	4.78E-03
NonpolarCountMW	-0.434
NPAT	6.33E-03
NPSA_percentage	4.66E-04
PAT	1.51E-05
PIAT	0.011
PolarCountMW	1.324
PSA	1.68E-04
PSA_percentage	-4.66E-04
SAS_HB_A_AREA	-2.17E-04
SAS_HB_D_AREA	-3.57E-04
SAS_POL_AREA	-9.24E-05
SPEC_HB_TOT	-0.018
SPEC_SAS_HB_A_AREA	-0.174
SPEC_SAS_HB_D_AREA	-0.201
SPEC_SAS_NONPOL_AREA	0.108
SPEC_SAS_POL_AREA	-0.108
SPEC_VDW_HB_A_AREA	-0.162
SPEC_VDW_HB_D_AREA	-0.216
SPEC_VDW_NONPOL_AREA	0.107
SPEC_VDW_POL_AREA	-0.107
VDW_HB_A_AREA	-1.17E-04
VDW_HB_D_AREA	-5.17E-04
VDW_POL_AREA	3.69E-05

9.7.4. HEMWat 20

Descriptor	Coefficient
ACDlogD6.5	0.031
ACDlogD7.4	0.024
ACDlogP	0.052
AverNegCharge_GH	-0.054
AverNegCharge_GM	0.498
AverPosCharge_GM	-3.45E-03
ChargeRang_GM	-0.191
ClogP	0.055
Constant	1.405
DipoleMom_GM	-2.04E-02
GClogP	0.042
HAROM	6.95E-03
HBA	-3.18E-03
HBA_nonlipinski	6.30E-03
HBA_Raevsky	-4.46E-03
HBA_SELMA	-9.43E-03
HBAmax	-9.13E-02
HBAsum	-0.013
HBD	-0.016
HBD_nonlipinski	-0.016
HBD_Raevsky	-0.016
HBD_SELMA	-0.026
HBDsum	8.30E-03
HBsum	1.66E-03
HBsumTotal	-9.23E-03
MAX_NEGCharge_GM	0.459
MAX_POSCharge_GM	-0.169
MaxRing1	0.015
MM_FHADSA	2.67E-04
MM_FHASA	1.56E-03
MM_FHDSA	-0.07
MM_HACA	-1.26E-03
MM_HADCA	1.56E-03
MM_HADSA	-2.86E-03
MM_HASA	9.17E-03
MM_HDCA	-0.399
MM_HDSA	-5.547
MM_MAX_NEG	0.085
MM_MAX_POS	-1.50E-03
MM_QMIN	0.085
MM_Qne_G_MEAN	-0.289
MM_Qneg_VAR	-0.173
MM_QO	6.94E-04
MM_QON	3.80E-03

MM_Qpos_MEAN	0.207
MM_Qpos_VAR	-0.067
MM_SAS_EP_N_MEAN	0.017
MM_SAS_EP_N_VAR	-0.024
MM_SAS_EP_NSUM	0.037
MM_SAS_EP_P_MEAN	-0.013
MM_SAS_EP_P_VAR	-0.013
MM_VDW_EP_N_MEAN	7.95E-03
MM_VDW_EP_N_VAR	-7.63E-03
MM_VDW_EP_P_MEAN	-3.72E-03
MM_VDW_EP_P_VAR	-1.05E-03
MWNPat	1.22E-03
MWPat	5.89E-04
MWSHDA	8.12E-04
NNlogP	0.04
NonpolarCount	5.14E-03
NonpolarCountMW	-1.271
NPAT	6.57E-03
NPSA_percentage	-1.73E-04
OxygenCount	-4.75E-04
PAT	-2.06E-03
PolarCount	9.80E-03
PolarCountMW	2.217
PSA	1.32E-04
PSA_percentage	1.74E-04
SAS_HB_A_AREA	-2.40E-04
SAS_HB_D_AREA	-7.01E-04
SAS_POL_AREA	-2.21E-04
SPEC_HB_TOT	0.012
SPEC_SAS_HB_A_AREA	-0.063
SPEC_SAS_HB_D_AREA	-0.32
SPEC_SAS_NONPOL_AREA	0.078
SPEC_SAS_POL_AREA	-0.078
SPEC_VDW_HB_A_AREA	-0.028
SPEC_VDW_HB_D_AREA	-0.361
SPEC_VDW_NONPOL_AREA	0.032
SPEC_VDW_POL_AREA	-0.032
VDW_HB_A_AREA	-1.51E-04
VDW_HB_D_AREA	-1.36E-03
VDW_POL_AREA	-1.16E-04

9.7.5. HEMWat 22

Descriptor	Coefficient
ACDlogD6.5	0.038
ACDlogP	0.047
ClogP	0.057
Constant	1.334
GClogP	0.031
HBD_SELMA	-0.065
MM_FHADSA	5.22 x 10-3
MM_FHASA	0.010
MM_HACA	-0.117
MM_HADCA	-0.092
MM_HADSA	-0.420
MM_HASA	-0.521
MM_HDCA	-1.076
MM_HDSA	-12.188
MM_Qneg_VAR	-0.054
MM_SAS_EP_N_MEAN	0.020
MM_SAS_EP_P_MEAN	-0.017
MM_SAS_EP_P_VAR	-0.024
MM_VDW_EP_N_MEAN	0.010
MM_VDW_EP_P_MEAN	-6.21E-03
MM_VDW_EP_P_VAR	-7.80E-03
MWSHDA	-5.77×10^{-4}
NNlogP	0.032
NonpolarCountMW	-2.568
NPSA_percentage	-3.45E-05
PolarCountMW	4.246
PSA	-1.28×10^{-3}
PSA_percentage	0.00
SAS_HB_A_AREA	-2.00
SAS_HB_D_AREA	-2.06E-03
SAS_POL_AREA	-9.64E-04
SPEC_HB_TOT	-0.068
SPEC_SAS_HB_A_AREA	-0.057
SPEC_SAS_HB_D_AREA	-0.810
SPEC_SAS_NONPOL_AREA	0.184
SPEC_SAS_POL_AREA	-0.184
SPEC_VDW_HB_A_AREA	-0.106
SPEC_VDW_HB_D_AREA	-0.984
SPEC_VDW_NONPOL_AREA	0.088
SPEC_VDW_POL_AREA	-0.088
VDW_POL_AREA	-1.56E-03

9.7.6. HEMWat 26

Descriptor	Coefficient
ACDlogP	0.083
ClogP	0.100
HBD_nonlipi	-0.016
HBD	-0.016
HBsum	-0.014
HBD_SELMA	-0.039
HBD_Raevsky	-0.016
HBDsum	0.012
HBsumTotal	-0.021
PSA	-0.001
SAS_HB_A_AREA	-0.002
SAS_HB_D_AREA	-0.001
SAS_POL_AREA	-0.001
SPEC_HB_TOT	0.129
SPEC_SAS_HB_A	-0.431
SPEC_SAS_HB_D	-0.341
SPEC_SAS_POL	-0.138
SPEC_VDW_HB_A	-0.463
SPEC_VDW_HB_D	-0.259
SPEC_VDW_POL	-0.054
VDW_HB_A_AREA	-0.003
VDW_HB_D_AREA	-0.003
VDW_POL_AREA	-0.002
PSApercent	0.001
NPSApercen	-0.001
SPEC_SAS_NON	0.138
SPEC_VDW_NON	0.054
ChargeRang	0.182
DipoleMomG	-0.053
MAX_NEGChar	-0.321
MM_FHASA	-0.009
MM_HACA	-0.214
MM_HADCA	-0.077
MM_HADSA	0.156
MM_HASA	-2.217
MM_HDCA	-0.194
MM_HDSA	0.302
MM_Qneg_VAR	0.784
MM_VDW_EP_NME	0.017
MWNPat	0.003
MWSHDA	0.000
OHCount	-0.084
MindistDA	-0.055
Constant	0.003

9.8. The combination of descriptors used to obtain the best performing QSAR models generated using MLR.

HEMWat system number	Combination of descriptors used to produce the model
8	Top 20 most significant descriptors selected using PLS (listed in section 9.5.1)
14	AZ Top14 descriptors (section 9.2)
17	AZ Top14 descriptors (section 9.2)
20	Five Abraham and Top 20 most significant descriptors selected using PLS (listed in section 9.6.4)
22	Five Abraham and Top 20 most significant descriptors selected using PLS (listed in section 9.6.5)
26	Top 20 most significant descriptors selected using PLS (listed in section 9.5.6)

9.9. Excel vs Simca prediction for the four test compounds. The two predictions that differed are highlighted in yellow.

Excel prediction	8	14	17	20	22	26
Benzoquinone	0.45	0.03	-1.23	-1.41	-1.73	-2.34
Biphenyl	2.78	2.04	1.18	0.87	0.58	-0.28
Quinine	1.90	0.68	-0.12	-0.80	-0.87	-1.41
Tolbutamide	1.77	0.13	-1.18	-2.01	-2.13	-1.99

SIMCA prediction	8	14	17	20	22	26
Benzoquinone	0.45	0.03	-1.23	-1.41	-1.73	-2.34
Biphenyl	2.78	2.04	1.18	0.87	0.58	-0.27
Quinine	1.90	0.68	-0.12	-0.80	-0.87	-1.40
Tolbutamide	1.77	0.13	-1.18	-2.01	-2.13	-1.99