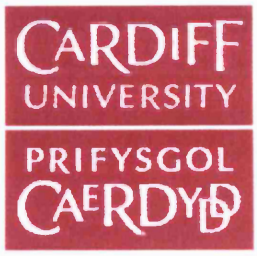


Molecular Surface Area Measures of Polarity and Hydrogen Bonding For QSAR

Robert Alun Saunders

Submitted for the degree of Ph.D.

July 2004

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Acknowledgments

First and foremost I would like to thank my supervisor Jamie Platts for giving me an interesting and challenging PhD, providing me with help and guidance throughout and allowing me to pursue me own ideas.

I would also like to thank all of the crew of lab 1.95 Arturo, Ed, Gareth, Jon, Marco, Nic, Rudy and Rajinder. I would also like to extend a special thank you to Farah and Olivier who provided me with help and entertainment from the beginning to the end.

I would also like to thank all of my friend outside chemistry Blan, Carl, Claire, Dean, Harry, Helen, Matt, Mike, Paul, Sam, Steve and Shippy whom without my sanity would not still be intact.

I would also like to thank Jenny for her constant love and support during the course of my PhD.

Finally I would like to say an extra special thank you my parents and my sister for all of their support (both emotional and financial) and encouragement during my university Career, without which I know I would never have got this far.

Abstract

Modifications were made to the traditional PSA descriptor by decoupling it into its H-bond acidic and basic components. The PSA based descriptors were also scaled according to the known hydrogen bonding characteristics of common functional groups to make them more realistic measures of a molecules hydrogen bonding capacity. Three other surface area descriptors total surface area, total halogen atom surface area and total aromatic carbon surface area were also defined.

Various routes to the calculation of these descriptors were explored and it was concluded the best descriptors were those obtained from a single structure generated using the semi empirical-method AM1. It was also shown that descriptors obtained from a vdw surface were more suitable than those obtained from solvent accessible surface area.

The scaled PSA descriptors were initially tested against octanol-water, chloroform-water, and cyclohexane-water partition coefficients of 110 organic and drug-like molecules. All of the models produced were seen to be statistically accurate and followed known characteristics of the partition coefficients considered.

The scaled PSA descriptors were then applied successfully to a number of important biological processes such as cellular uptake and intestinal absorption; models were also produced for important industrial processes such as Fluorophilicity and CMC. The surface area descriptors were also seen to be equally capable of modelling inorganic molecules and excellent models were produced for octanol-water and chloroform-water partitions for a number of platinum containing drugs.

Publications

Chapter 3 and 4

Scaled polar surface area descriptors: development and application to three sets of partition coefficients
Saunders RA, Platts JA
NEW JOURNAL OF CHEMISTRY 28 (1): 166-172 2004

Chapter 6

Statistical and theoretical studies of fluorophilicity
Huque FTT, Jones K, Saunders RA, Platts JA
JOURNAL OF FLUORINE CHEMISTRY 115 (2): 119-128 JUN 28 2002

Prediction of fluorophilicity of organic and transition metal compounds using molecular surface areas
Daniels S, Saunders RA, Platts JA
JOURNAL OF FLUORINE CHEMISTRY (IN PRESS)

Linear free energy relationship analysis of the solubility of solids in supercritical CO2
Saunders RA, Platts JA
JOURNAL OF PHYSICAL ORGANIC CHEMISTRY 14 (9): 612-617 SEP 2001

Correlation and prediction of critical micelle concentration using polar surface area and LFER methods
Saunders RA, Platts JA
JOURNAL OF PHYSICAL ORGANIC CHEMISTRY 17 (5): 431-438 MAY 2004

Glossary

Abbreviations

AM1	Austin Method one
AMBER	Assisted Model Building and Energy Refinement
BBB	Blood Brain Barrier
Caco-2	Human colon carcinoma cells
CMC	Critical Micelle Concentration
CNS	Central Nervous System
CoMFA	Comparative Molecular Field Analysis
LFER	Linear Free Energy Relationship
LSER	Linear Solvation Energy Relationships
MX	Cuticular Polymer Matrix Membrane
PAH	Polycyclic Aromatic Hydrocarbons
PAMPA	Parallel Artificial Membrane Permeability Assay
QM	Quantum Mechanical
QSAR	Quantitative Structure Activity Relationship
QSPR	Quantitative Structure Property Relationship
RBR	Relative Biological Response
SASA	Solvent Accessible Surface Area
SMILES	Simplified Molecular Input Line Entry Specification
vdw	van der Waals
WHIM	Weighted Holistic Invariant Molecular
CODESSA	Comprehensive Descriptors for Structural and Statistical Analysis

Statistics

MLRA	Multiple Linear Regression Analysis
N	Number of data points used in model
PLS	Partial Least Squares
R^2	Multiple Correlation Co-efficient
R^2_{CV}	Cross validated R^2
RMSE	Route Mean Square Error
Sd	Standard Deviation

Descriptors

Surface Area descriptors

ASA_U	Unscaled Hydrogen Bond Acidic Surface Area
ASA_S	Scaled Hydrogen Bond Acidic Surface Area
BenSA	Benzene Surface Area
BSA_U	Unscaled Hydrogen Bond Acidic Surface Area
BSA_S	Scaled Hydrogen Bond Acidic Surface Area
ClSA	Chlorine Surface Area
FSA	Fluorine Surface Area
HalSA	Halogen Surface Area
MetalSA	Metal Surface Area
O^-SA	Anionic Oxygen Surface Area
PSA_U	Unscaled Polar Surface Area
PSA_S	Scaled Polar Surface Area
TSA	Total Surface Area
PSA_d	Dynamic Polar Surface Area

Abraham Descriptors

E	Excess Molar Refraction
S	Polarity/Polarisability
A	Hydrogen bond Acidity
B	hydrogen Bond Basicity
V	McGowan's Characteristic Volume

Properties

AbsT	Arcsin Intestinal Absorption
K_{cc}	Uptake by Chara Ceratophylla Cells
K_{nit}	Uptake by Nitella Cells
L_{Blood}	Gas/Blood Partition
L_{Brain}	Gas/Brain Tissue Partition
L_{Fat}	Gas/Fat Partition
L_{Heart}	Gas/Heart Tissue Partition
L_{Kidney}	Gas/Kidney Tissue Partition
L_{liver}	Gas/Liver Tissue Partition
L_{Lung}	Gas/Lung Tissue Partition
M_{uscle}	Gas/Muscle Tissue Partition
$logK_{MXa}$	Air/Cuticular Polymer Matrix Membrane Partition
$logK_{MXw}$	Sater/Cuticular Polymer Matrix Membrane Partition
logS	log Water solubility
L_{Oil}	Gas/Oil Partition
L_{Plasma}	Gas/Plasma Partition
L_{water}	Gas/Water Partition
P_{app}	Apparent Intestinal Permeability
P_{CHCl}	Water/Chloroform partition
P_{cyc}	Water/Cyclohexane partition
P_o	PAMPA permeability
P_{oct}	Water/octanol partition
S_{CO2}	Solubility in Super Critical CO_2

INDEX

Chapter One Introduction

1.1 Introduction	1
1.2.1 QSAR	1
1.2.2 History of QSAR	2
1.2.3 QSAR Descriptors -Electronic Effects	3
1.2.4 QSAR Descriptors Steric Effects	4
1.2.5 Multiparameter QSAR	5
1.2.6 3D QSAR	5
1.2.7 Other QSAR Descriptors	7
1.3 Solvation	8
1.3.2 Octanol-Water Partition	8
1.3.2 Group Contribution Methods	9
1.3.3 Atomic Contribution Methods	10
1.3.4 Molecular Methods	10
1.3.5 Linear Solvation Energy Relationships (LSER)	11
1.3.6 Molecular Surfaces	13
1.4 Polar Surface Area	16
1.4.1 History of PSA	16
1.4.2 PSA and Surface Area Type	18
1.4.3 Dynamic PSA	19
1.4.4 Rapid Calculation of PSA	21
1.4.5 PSA and Additional Descriptors	22
1.4.6 What does PSA represent?	23
1.4.7 PSA and Hydrogen Bonding Strength	24
1.4.8 Charged Partial Surface Areas	25
1.5 Conclusions	26
1.6 References	27

Chapter 2. Statistical Methods 31

2.1 Linear Regression 31
2.2 Multiple Linear Regression 33
2.3 Root Mean Square Error 34
2.4 F-ratio 35
2.5 Cross-validated R^2 and Training and Test Sets 35
2.6 Interpretation of MLRA Equations 36
2.7 Stepwise Regression 37
2.8 Outliers 37
2.9 Limitations to MLRA and Alternatives 38
2.10 Software 40
2.11 References 40

Chapter 3. Model development

3.1 Generating Descriptors 41

3.1.1 MOLVOL 41

3.1.2 Descriptors 47

3.1.3 Scaling Factors 50

3.1.3.1 Averaging of Abraham Scales 51

3.1.3.2.1 Regression of Abraham Scales 54

3.1.3.2.2 Results ASA_S 54

3.1.3.2.3 Results BSA_S 55

3.1.3.3 Scaling Factors Discussion 58

3.1.4 Fragments 60

3.1.5 Output 64

3.1.6 Automation 65

3.2 Generating Structures

3.2.1 CORINA 66

3.2.2.1 Optimisation Methods 67

3.2.2.2 Geometry Optimisation Conclusions 75

3.2.3 Conformational Flexibility and its Effects on PSA 76

3.2.4 Encoding 3D Information 79

3.2.4.1 Solvent Accessible Surface Area 81

3.2.4.2 Expanded van der Waals Radii 82

3.2.4.3 GETAREA 84

3.2.5 Conformations and SASA Conclusions 86

3.3 References 88

Chapter 4. Partition Models

4.1 Data 90

4.2.1 Descriptors and Scaling Factors 90

4.2.2 Results and Discussion 91

4.2.3 Significance of Descriptors 97

4.2.3 Predictive Accuracy 97

4.3 Geometry Optimisation 99

4.4 Effects of Conformational Change Within Models 101

4.5 Solvent Accessible Descriptors Expanded Radii method 103

4.6 Complex molecules: Peptides Expanded radii method and Conformation changes 105

4.7 Solvent Accessible Descriptors GETAREA 108

4.8 PSA Descriptors and 3D Information 110

4.9 Model Development Conclusions 110

4.10.1 Cisplatin Complexes 114

4.10.2 Introduction 114

4.10.3 Methods 115

4.10.4.1 $logP_{oct}$ Results 116

4.10.4.2 $logP_{chcl}$ Results 120

4.10.5 Combining Organic and Inorganic Datasets 121

4.10.6 Coefficient Analysis 124

4.10.7 Platinum Complex Conclusions 127

4.11 Future Work 128

4.12 References 129

Chapter 5 Biological Properties

5.1 The uptake of Volatile organic compounds into the cuticular matrix of plants

5.1.1 Introduction **130**

5.1.2 Methods **133**

5.1.3 Results and Discussion **133**

5.2 Partition into Biological liquids and tissues of vapours and biological liquids

5.2.1 Introduction **136**

5.2.2 Method **138**

5.2.3 Results and Discussion **138**

5.3 Blood brain barrier

5.3.1 Introduction **147**

5.3.2 Methods **149**

5.3.3 Results **148**

5.3.3.1 Dataset I **150**

5.3.3.2 Dataset II **151**

5.3.3.3 Dataset III **153**

5.3.4 Discussion **155**

5.4 Intestinal Absorption

5.4.1 Introduction **159**

5.4.2. Methods **160**

5.4.3.1 Results Dataset I **161**

5.4.3.2 Results Dataset II **162**

5.4.3.3 Results Dataset III **163**

5.4.3.4 Passive Permeability Introduction **165**

5.4.3.5 Passive Permeability Results **166**

5.4.4 Discussion **167**

5.5 Cell Permeation

5.5.1 Introduction **170**

5.5.2 Methods **171**

5.5.3.1 Results Chara ceratophylla Cells $logK_{cc}$ **171**

5.5.3.2 Results Nitella Cells $logK_{nit}$ **172**

5.5.3.3 Results Dead Nitella Cells $logK_{nit}$ **173**

5.5.3.4 Results Combined Chara ceratophylla and Nitella Cells $logK_{gen}$ **173**

5.5.4 Discussion **174**

5.6 Conclusions

5.6.1 Conclusions -Uptake of volatile organic compounds by plants **175**

5.6.2 Conclusions - Partition into Biological liquids and tissues of vapours and biological liquids **175**

5.6.3 Conclusions - Blood Brain Barrier **176**

5.6.4 Conclusions - Intestinal Absorption **176**

5.6.5 Conclusion Cellular Uptake **177**

5.7 Future work **178**

5.8 References **179**

Chapter 6 – Industrial properties and green solvents

6.1 Fluorophilicity

6.1.1 Introduction **181**

6.1.2 Method **183**

6.1.3 Results and Discussion **185**

Chapter 6.2 Solubility in Supercritical Carbon dioxide

6.2.1 Introduction **190**

6.2.2 Methods **192**

6.2.3 Results **193**

6.3 Critical Micelle Concentration

6.3.1 Introduction **199**

6.3.2 Method **202**

6.3.3 Results **200**

6.3.3.1 Dataset 1: Non ionic **203**

6.3.3.2 Dataset 2: Ionic Surfactants **205**

6.3.3.3 Combined Dataset **207**

6.3.3.4 Dataset 3: Structurally Diverse Drug Molecules **208**

6.3.4 Discussion **209**

6.4 Conclusions

6.4.1 Conclusions: Fluorophilicity **212**

6.4.2 Conclusions: Supercritical CO_2 **212**

6.4.3 Conclusions: CMC **213**

6.5 Future Work **213**

6.6 References **214**

7.1. Closing Remarks 217

7.2 References 218

Chapter 1. Introduction

1.1 Introduction

The ability to predict *a priori* the properties of molecules, especially drugs and other bio-active molecules has led to a great deal of research interest,[1] especially within the pharmaceutical and agrochemical industries. Prediction of biological activity is one of the major goals of such studies. However, the capacity of a proposed drug to act against a particular target is irrelevant if the molecule cannot reach the target. Hence the prediction of properties such as lipophilicity, aqueous solubility, bioavailability, and CNS penetration at the beginning of the product design process allows early identification and removal of candidate molecules with unsuitable characteristics. The independence of a predictive method from experimentally observed information also allows prediction of chemical properties without the need for synthesis. However the use of experimental observed data in predicting properties is still advantageous when a difficult to measure property can be related to a property that is easily measured. For example water solubility (difficult) and Skin permeability (difficult) can be calculated from properties such as water-octanol partition, melting or molecular weight.[2]

1.2.1 QSAR

A quantitative structure activity relationship (QSAR) is a method by which numerical properties derived from molecular structure is mathematically related to its activity. At its most general, a QSAR equation takes the form.

$$\text{Activity} = f(X_1, X_2, \ldots . X_B) \qquad (1.1)$$

Activity is a function of molecular properties or descriptors $X_1 \ldots X_B$ which encode important chemical information about the molecule. It is therefore important that a QSAR equation contains descriptors that encompass the properties that have influence on the activity.

The term QSPR (quantitative structure property relationship) is used to refer to a relation where the property of interest is not a biological activity. Through analysis of the form of

QSAR/QSPR equations it is possible to interpret which structural features are beneficial to activity and which are a hindrance; this information can be used to guide the design process of pharmaceuticals and help make compounds with more appropriate activities or other properties.

Where a relation is correlating Gibbs free energy related variable (e.g. equilibrium constant) and descriptors are related to well-defined interactions, the term linear free energy relationship (LFER) is often used.

1.2.2 History of QSAR

One of the earliest examples of a SAR was that proposed by Crum Brown and Fraiser[3] in 1868. They noted that curare-like properties of a series of quarternised strychnines was dependent upon the quaternised group and therefore proposed that the physiological activity was a function of the structure of the molecule.

In 1869 Richardson reported the relationship between toxicity of simple ethers and alcohols and their water solubility.[4] Similarly Overton[5] and Meyer,[6] working independently, observed a relationship between aqueous narcosis of tadpoles and partition coefficients. A partition coefficient (P) is a measure of the affinity of a molecule for a solvent phase (in this case olive oil) versus that for water, calculated as the equilibrium ratio of concentration of solute in the solvent phase to that in the water phase. Overton's interpretation was that the narcotic effect was due to physical changes caused by the dissolution of the drug in the lipid component of cells. Although it was not until 1920 that Meyer and Gottlieb-Billlroth proposed the relationship in a quantitative formula,[7] this represents the first reported QSAR.

$$C_{nar} * P_{oil} = \text{Constant.} \qquad (1.2)$$

Where C_{nar} is the concentration required to induce narcosis and P_{oil} is the olive oil/water partition. Gottlieb-Billroth also showed a relation between aqueous narcosis and water oleyl alcohol partition, and proposed a QSAR for gaseous narcosis.

In 1933 Collander and Barland[8] offered the first real examination of partition coefficients. Collander demonstrated that the partition coefficients of two similar water solvent systems (I and II) could be related thus.

$$LogP_{II} = a.logP_I + b \qquad (1.3)$$

Where a and b are constants. Collander[9,10] stated that the relationship would only be valid where the two solvents had similar chemical properties. The ideas of Collander were expanded by Leo and Hansch,[11] who also suggested octan-1-ol as a standard solvent, as the water/octanol partition was a better measure of lipopilicity (the $logP_{oct}$ partition is discussed at greater length in 1.3).

1.2.3 QSAR Descriptors -Electronic Effects

Hammett[12-14] undertook a systematic study of a series of benzene derivatives to establish a set of descriptors which would encode electronic properties, and could be applied to other solvents and used for prediction of activity and other properties. Hammett defined a descriptor σ, referred to as the Hammett substituent constant. This descriptor is a measure of the electron donating or withdrawing strengths of the substituents. Values of σ are determined by comparing the equilibrium constants for the ionisation of substituted benzoic acids to the ionisation constants of unsubstituted benzoic acid.

$$\sigma = \log[K_x] - \log[K_H] \qquad (1.4)$$

Where K_X and K_H are equilibrium constants of ionisation in a given system for benzoic acid, and X substituted benzoic acid respectively. Benzoic acid by definition has σ equal to 0; positive values of σ are given by electron withdrawing subtituents due to the stabilising effect on the anion, while electron donating subtituents give negative values for σ. The values of σ are usually seen to be different for the same subsitituent if it is located in the meta or para position, an effect which is attributed to the enhanced resonance effects at the para position. Using the σ descriptor Hammett derived the following equation.

$$\text{Log } K_X = \rho\,\sigma + \log K_H \qquad (1.5)$$

where ρ is the slope of the line of best fit, and is a measure of how sensitive the given system is to the electronic effects of the substituents.

Hammett σ values are still widely regarded as one of the most reliable and general means to assess the electronic effects of substituents upon a reaction; a large number of σ values have been reported, and Hammett type relationships have been applied to a wide variety of physiochemical studies.[15-17]

1.2.4 QSAR Descriptors Steric Effects

The work of Hammett was expanded upon by Taft[18] in the 1950s. Taft proposed a descriptor E_S that would account for the steric effects of substituents upon activity. Taft compared the effects of subtituents upon the hydrolysis of esters in acidic conditions.

$$E_s = \left(\frac{K^x}{K^o}\right)_{acid} \qquad (1.6)$$

Where K^O represents the methyl derivative and K^X is the X-substituted compound. The hydrolysis was performed under acidic conditions, as the hydrolysis of esters under acidic conditions is dependent only upon steric effects where hydrolysis under basic condition is dependent on steric and electronic effects. This dependence and independence of electronic effects allowed Taft to define a further electronic descriptor σ_i.

The acid disassociation constant of 4-substituted bicyclo[2.2.2]octane carboxylic acid in 50% aqueous ethanol was used by Roberts and Moreland[19] as a more direct method of determining σ_i.

$$\sigma_i = 0.606\left(\frac{K_z^x}{K_a^H}\right) \qquad (1.7)$$

Where K^X_a and K^H_a are the acid dissociation constants for the substituted and unsubsitituted compounds respectively. The value of 0.606 is a scaling factor to place the values of σ_i on the scale proposed by Taft.

1.2.5 Multiparameter QSAR

Hansch[20,21] realised that in order to fully describe activity or other properties, a single descriptor would be insufficient as any property is the result of a combination of different factors, and that multiparameter QSAR was necessary to account for all these factors. The first use of a multiparameter QSAR equation by Hansch in 1964 is considered by many to be the origin of modern QSAR. Hansch stated the following equation for the prediction of relative biological response (RBR).

$$LogRBR = c - a\pi^2 + b\pi + \rho\sigma + E_S \qquad (1.8)$$

Where σ is Hammett's electronic parameter, E_S is Taft's steric parameter and π is the difference in logP for substituted and unsubstituted benzene as shown in equation 1.9.

$$\pi = \log P_X - \log P_H \qquad (1.9)$$

1.2.6 3D QSAR

The boundaries of QSAR were further expanded by the introduction of 3D QSAR. 3D QSAR methods tend to treat the molecule as a whole rather than a collection of substituents, offering the ability to account for conformational flexibility, ring and cage formation, and intramolecular interactions.

The first published 3D QSAR method was the Comparative Molecular Field Analysis (CoMFA) model of Cramer in 1988.[22] In CoMFA, a set of conformations is generated, one for each molecule in the set. This conformation is presumed to be the conformation of the active structure. These conformations are then overlaid against each other in the proposed binding mode. The molecular field surrounding each conformation is then calculated by placing appropriate probe groups at points (usually a distance of 2Å

between points) on a regular lattice that encompasses the molecules. The type of probe used is dependent of the type of interaction.

The results of this analysis can be represented as a matrix S, in which each row corresponds to one of the molecules and the columns are energy levels at the grid points. If there are N points in the grid and P probe groups are used there will be N x P such columns. The table is completed by adding a column that contains the relevant activity of the molecule. A correlation between biological activity and the field values is then determined. The general form of the equation that is desired is thus

$$Activity = R + \sum_{i=1}^{N} \sum_{j=1}^{P} C_{ij} S_{ij} \qquad (1.10)$$

Where C_{ij} is the coefficient for the column in the matrix that corresponds to placing probe group J at grid point I. Owing to the large number of descriptors generated from CoMFA partial least squares (PLS) analysis (see 2.9) is required.

PLS analysis generates a coefficient for each column in the data table. This coefficient indicates the significance of each grid points to the activity. This information is most usefully represented as a 3D surface that connects point which have the same coefficients. These diagrams are then used to identify regions where (for example) changing the steric bulk would increase or decrease the activity/binding. As with any QSAR, prediction of the activities of molecules not included in the analysis is possible by calculation the fields of these molecules and inserting these values in to the equation obtained from the regression.

CoMFA still remains one of the most popular 3D methods in the drug design with hundreds of applications in the field of ligand protein interactions.[23,24]

1.2.7 Other QSAR Descriptors

The range of descriptors that have been proposed for use in QSPR methods to date is overwhelming. QSPR descriptors can broadly be divided into two categories: observed and theoretical/calculated descriptors. Observed descriptors are properties such as boiling and melting point, spectroscopic shifts and acidities/basicities.

Theoretical descriptors are more widespread and range from simple counts of hydrogen bond donors and acceptors to more complex descriptors such as HOMO and LUMO energies, electrostatic potentials and partial atomic charges.[25,26] Some models have also proposed simple indicator variables as descriptors to account for specific structural properties that are important to the solvation property but not encoded by any available QSPR. The popularity of theoretical descriptors has been driven largely by the increased use of high throughput screening and combinatorial where the large numbers of compounds involved (often measured in millions) make measurement of properties by experimental methods expensive and time consuming. The enormous number of analogues produced by combinatorial chemistry also means that any predictive method must be rapid if it is to be used successfully as a virtual screening tool. Thus theoretical descriptors that do not require measurement of experimental data, such as partition coefficients or melting points, offer a considerable advantage. A further advantage to theoretical descriptors is they offer the ability to predict properties of a molecule without the need for synthesis.

The huge amount of descriptors that are available to describe the properties of molecules can make the process of selecting descriptors for a QSPR difficult. One approach to this problem is found in QSPR methods which select an appropriate subset from larger pools of up to several hundred possible descriptors, choosing a new subset for each property of interest.[27,28] In this way, it is possible to find correlations that could otherwise have been missed. This approach has been taken by Jurs and Kattritsky who developed the program CODESSA.[29] For a model of aqueous solubility for 258 liquids made using CODESSA, Jurs selected nine descriptors from a pool of 157 descriptors.[30] While these methods are predictively accurate and can find correlations that are missed by other methods they are also usually quite difficult to interpret.

1.3 Solvation

The conclusions drawn by Richardson,[4] Overton[5] and Meyer[6] as discussed in 1.2.2 have been widely accepted and the influence of molecular properties such as solubility and partition coefficients upon observed biological activity have been studied at great length.[31-37] As a result of this many specialist QSAR and QSPR methods have been formulated for the prediction of solvation properties.[38]

1.3.1 Octanol-Water Partition

Since its first use by Collander[10] octanol-water partition has become the standard measure of lipophilicity in QSAR. The ability of this system to represent lipophilicity has been attributed to its chemical similarities to the partition between aqueous and biophases.[33,39] Octanol, with a polar head and flexible non polar tail has hydrogen bonding capabilities and amphiphilicity characteristics similar to those of the phospholipids and proteins that make up biological membranes. The octanol water partition coefficient is defined thus.

$$\log P_{oct} = \log \frac{[octanol]}{[water]} \tag{1.11}$$

Where [octanol] and [water] are the concentrations of the solute in the octanol and water phases respectively. The term is expressed as a logarithm because of the wide range covered, as much as 8-10 orders of magnitude. For a molecule partitioning between water and octanol the molecule is said to be lipophilic if $P > 1$ and hydrophobic where $P < 1$, with the vast majority of used drugs having $\log P_{oct}$ values in a relatively tight range, somewhere between 1 and 5.[40]

The popularity of $\log P_{oct}$ within QSAR studies can also be attributed to the large amounts of reliable experimental data that is available,[41,42] this is due to most compounds of interest having $\log P_{oct}$ values that are in a range that can easily be experimentally observed using standard (e.g.) shake flask methods.

Other partitions have been proposed, although the results of many studies have shown $\log P_{oct}$ to give the best correlation with biological activity.[43] While logP has no close

rival, the use of water/cyclohexane partition has been proposed and used with some success.[44]

Numerous methods have been proposed for the prediction of $logP_{oct}$ many of which have been rigorously review and compared.[40,45-50]

1.3.2 Group Contribution Methods

One common method of predicting logP values is through the use of a "group contribution approach", in this method molecules are broken down into a series of predefined fragments and their corresponding contributions are summed to obtain a final model. One of the earliest group contribution methods was that of Rekker,[51,52] who proposed a set of 136 fragments and then calculated their contribution to $logP_{oct}$ via multiple linear regression analysis (MLRA) against observed values of logP. Rekker also included 10 correction factors to account for the fact that a molecule's properties were more than collection of fragments. In order to overcome the need for correction factors Suzuki and Kudo[53] produced a model of logP that used 494 fragments.

A similar method to that of Rekker was employed by Hansch and Leo, who developed the program CLOGP.[41,48,54,55] The method of generating fragments used by Hansch and Leo differed to that of Rekker in that they were derived from experimental $logP_{oct}$ values for a small set of simple compounds. The initial CLOGP program used 200 fragments and 20 correction factors[41] although numerous other correction factors and fragments have been subsequently added to the CLOGP which have substantially improved the model.[56,57]

Bodor[40] used the methods of Rekker and CLOGP to model $logP_{oct}$ for 145 molecules many of which were peptides, nucleotides, druglike and halogenated aromatic molecules. The results showed the method of Rekker to be inferior to that of CLOGP giving an R^2 of 0.757 compared to 0.934 yielded by CLOGP, although when the models were created from a subset of 101 molecules (exclusion of all the halogenated molecules from the original dataset) the difference between the two methods was smaller with R^2 values of 0.887 and 0.844 respectively.

1.3.3 Atomic Contribution Methods

Another popular method for the calculation of $logP_{oct}$ has been through the use of "Atomic contribution methods". These methods are similar to group contribution methods except they try to predict $logP_{oct}$ using single atom contributions. This method has been noted as being problematic[58] as the problems that arises by treating a molecule as just a sum of fragments under the group contribution method is exacerbated when further broken down in to atomic contributions, atoms are defined in relation to their topology and environment within the molecule, for example in the AlogP method of Ghose and Crippen which initially used 110 descriptors[59,60] but was later extended by Viswanadhan (ALOGP)[61] to 120 descriptors, has 44 atom types defined for carbon and 10 different atom types defined for hydrogen.

Another method to overcome the problem that a molecule is more that the sum of its parts is to define numerous groups of atoms as a fragment via specific bonding pathways, as implemented in the method of Broto *et al*[62] where 222 descriptors were used many of which consisted of combinations of up to four atoms. Similarly to the group contribution method correction factors are also included in the descriptors. Despite these problems, atom contribution methods are of interest due to the relative ease of their computer implementation.

1.3.4 Molecular Methods

Through the use of quantum chemical modelling numerous descriptors have been defined[63] that more accurately represent the interaction of the solute with the surrounding solvent system and treat the molecule as a whole.

In 1981 Klopman and Iroff[64] used atomic charges obtained from molecular orbital calculations to estimate logP values for 61 simple organic molecules. Another and more widely used molecular method is that of Bodor.[65,66] This method uses geometric properties such as volume and surface area and electronic distribution parameters such as dipole moment and charge density obtained from AM1 calculations. Using this method 302 $logP_{oct}$ values were modelled with an R^2 of 0.98 and an Sd. (standard deviation) of

0.31. Sasaki *et al*[67] used surface tension, electrostatic potential and charge transfer, derived from structures optimised with molecular mechanic methods to calculate $\log P_{oct}$ for 63 compounds.

Another group of descriptors which contain information about the whole molecular structure are the Weighted Holistic Invariant Molecular (WHIM) indices.[68] The descriptors are calculated from 3D structures and weighted by atomic mass, van der Waals atomic electronegativities and geometric parameters. WHIM descriptors were applied to a dataset of 268 small molecules and using PLS analysis $\log P_{oct}$ values were reasonably well predicted (R^2=0.77, Sd= 0.66) although the use of WHIM descriptors is limited due to the computationally demanding need for quantum mechanical calculation.

The 3D QSAR method CoMFA which also treats the molecule as a whole has also been applied to the prediction of $\log P_{oct}$ although it has only been applied to small specific datasets such as furans and triazines.[69]

1.3.5 Linear Solvation Energy Relationships (LSER)

Linear solvation energy relation ships (LSER) represent a general and rigorous physiochemical treatment of solvation effects. These methods describe a large number of solvation effects with an equation that assumes solvation properties can be decomposed into cavity formation, dipolarity/polarisability and hydrogen bonding effects.[70-72]

One of the first rigorous LSER methods was that of Kamlet, Taft and Abraham[18,73-76] who developed a multiparameter QSPR for the prediction of transport properties of a given solute upon solvents.

$$\log Tr = c + d\delta + s\pi^*_1 + a\alpha_1 + b\beta_1 + d(\delta_H)^2 \qquad (1.12)$$

Where logTr is the transport property of a given solute, δ is an empirical solvent polarisability correction term, π^*_1 is the solvent dipolarity/polarisability descriptor α_1 and β_1 are solvent hydrogen bond acidity and basicity descriptors respectively, and (δ_H) is the Hilderbrand cohesive energy density, which is the energy needed to remove a molecule

from its nearest neighbour. Following on from this work, Kamlet produced an equation for the transport properties of solutes in a given solvent.[72,77,78]

$$\log SP = c + d\delta + s\pi^* + a\alpha + b\beta + vV \qquad (1.13)$$

LogSP is the log of a solubility property, δ is an empirical solute polarisability correction term, π^* is the solute dipolarity/ polarisabilty descriptor, α and β are solute hydrogen bond acidity and basicity descriptors respectively, and V is the solute volume.

The advantage to the solute parameter approach is that important physiochemical information that governs the properties of the solute are encoded into the descriptors, these properties are observed independently of any solvent. Hence the equation is far more general and can be applied to other solvation or biological properties.

Following on from the work of Taft and Kamlet and an approach that merits further discussion due to its applications within this study is the Linear Free Energy Relation (LFER) method of Abraham and co-workers.[79] In this method, solvation properties are expressed as linear combinations of molecular descriptors, according to equation 1.14.

$$\log SP = c + eE + sS + aA + bB + vV_x \qquad (1.14)$$

Where logSP is some solvation property; E is the excess molar refraction and is a measure of the dispersion interactions of π and n electron pairs; S is a joint polarity/polarisability term; A and B are hydrogen bond donor and acceptor strengths; and Vx is McGowan's characteristic molecular volume.[80] The LFER approach is sufficiently general to model many properties of interest and has been applied to solubility in water, organic solvents and to important biological properties such as blood-brain distribution, intestinal absorption and uptake into plants.[81]

Calculation of the necessary descriptors was traditionally a slow method, requiring manual data analysis and often experimental data input. However, a rapid, automated fragmentation method for their estimation[82] based upon simple atom and functional group definitions has been established. The values of descriptors calculated using this fragmental approach have been seen to be close to the value of descriptors obtained from

experimental results. The quality of these descriptors has been further verified through the successful modeling of numerous physiochemical properties.

The strength of the Abraham equation lies in the fact that the descriptors used in this method were carefully chosen to model specific interactions that are crucial to many solvation and biological properties, meaning that just five descriptors are applicable across a wide variety of solvents.

Using equation 1.14 Abraham showed excellent correlations for 613 molecules and $\log P_{oct}$ (R^2=0.994, Sd= 0.12).[83] The LFER equation is shown bellow.

$$\log P_{oct} = 0.088 + 0.562E + 1.054S + 0.034A - 3.460B + 3.81V \quad (1.15)$$
$$R^2=0.994 \; Sd=0.12$$

From analysis of the regression coefficients obtained from the equation 1.15 Abraham stated that the positive E term indicates octanol is able to interact with π and n electron pairs of the solute greater than water. The positive value of the S term shows that octanol is less polar/polarisable than water. The positive value for A is only small which indicates that water and octanol are similar in hydrogen bond basicity, where as the large negative B term indicates that water is much more hydrogen bond acidic that octanol. Finally the large V term indicates that large molecules will be more preferentially partitioned in to water. The conclusions drawn from the relatively large values of B and V are in fitting with conclusions from other studies that suggest $\log P_{oct}$ is governed largely by hydrogen bond basicity and molecular size.[84,85]

1.3.6 Molecular surfaces

Another method to encode 3D information into QSAR models and to treat a molecule as a whole is through the use of molecular surface areas. The use of molecular surface areas in the modelling of solvation properties has a long history. Langmuir[86] was the first to suggest the use of the molecular surface area in the estimation of solution free energies.

One of the earliest studies in which surface areas were used quantitatively to estimate solvation properties was that of Hermann in 1970.[87] Hermann calculated the surface area of a cavity that would need to be formed to accommodate a molecule in water. The cavity surface area was defined as the area traced out by the centre of water sized probe sphere rolled across the surface of the solute. The cavity surface area is effectively a measure of the number of solvent molecules that can be packed around a solute.

Solvent cavity surface areas were calculated for a series of hydrocarbons, and then correlated against the logarithm of their water solubility (logS), a linear relationship between solvent cavity size and logS was reported. Hermann noted that aromatic hydrocarbons were lower in their logS values than corresponding open chain analogues and did not fit on the same line due to the increased water solubility of the ring systems. Hence, two correlations were produced, one for purely aliphatic and one for aromatic hydrocarbons. Hermann's results showed as that as the surface area of the cavity is increased the solubility decreases.

Hermann noted that when applied to molecules containing polar functional groups, cavity surface area did not describe the modification to the water structure from the polar group. Assuming that functional groups and hydrocarbons contribute differently to solubility, Amidion[88] subdivided total cavity surface area in to hydrocarbon and functional group surface area (HYSA and FGSA respectively), in order that the approach could be applied to functionalised solutes. Amidion proposed the following relation.

$$\log S = C + C_1 HYSA + C_2 FGSA + C_3 IFG \qquad (1.16)$$

Where logS is the log of the solubility, C is the intercept and IFG is the functional group index, an indicator descriptor with a value of zero for hydrocarbons and 1 for monofuctional molecules, C_1, C_2 and C_3 are regression coefficients.

While surface areas were calculated in a manner similar to Hermann, Amidion refers to the surface area descriptor as a measure of the solvent accessible surface area (SASA) a nomenclature that is now used almost exclusively for this type of surface.

Amidion produced separate models for the aqueous solubility of the following monofunctional molecules: ethers, ketones, aldehydes, carboxylic acids, esters and olefins. The overall statistics for these models was excellent, with an average R-value of 0.991 and an average Sd of 0.167. Amidion showed that in these models the separation of total surface area was possibly unnecessary as both descriptors modelled similar effects so the following model was proposed with total solvent accesible surface area (TSASA).

$$LogS = C + C_1TSASA + C_2IFG \qquad (1.17)$$

Equation 1.17 was seen to model the effects of individual datasets with similar accuracy to equation 1.16. Amidion stated that for all the groups of compounds similar coefficient values of IFG and TSASA were given except for the olefins. For this reason the olefins were removed and all remaining groups were combined to make a dataset of 227 compounds, to which equation 1.17 was applied. The regression produced excellent correlation with $R^2 = 0.98$ and a Sd of 0.216. Throughout all model the IFG was seen to give a negative value, i.e. aiding solubility.

Similar partitioning of total SASA into polar and non polar regions was performed by Dunn.[89] Dunn calculated these surfaces using the same methodology as Herrman but using the algorithm of Lee and Richards.[90] This algorithm was originally designed to interpret the conformational effects upon the surface area of peptides, proteins and biopolymers but has meet with considerable success and acceptance as a method of generating surface descriptors in QSAR and QSPR studies.

Using principle component analysis to model the partition coefficients of octanol, ether, chloroform, benzene, carbon tetrachloride and hexane for 50 solutes, Dunn showed that the non polar, or isotropic surface area accounted for 80% of the partitioning data, while the functional group surface area accounted for 15% of the variance. Dunn interpreted that the 80% of variance accounted for by the isotropic surface was representing the "non specific" interactions of the solute and water.

1.4 Polar Surface Area

1.4.1 History of PSA

While the work of Amidion[88] and Dunn[89] showed the surface area of functional groups to be less influential than total surface area upon solubilities and partition values, McCracken and Lipkowitz,[91] in a study of the structural activity relationships of benzothiazole and benzimidazole anthelmintics, showed a clear relationship between observed anthelmintic activity and the polar surface area (PSA). Both PSA and percentage PSA were seen to correlate strongly and negatively with the log of the dosage of anthelmintic needed to produce a response. McCracken said this result could be interpreted in two ways: either that drugs with higher PSA are more soluble in polar media such as water, and may be better transported, or that the active site its self is very polar and drugs with higher PSA are better able to bind with the active site to elicit an effect. Distinguishing between the two effects was difficult, and it could not be determined without ambiguity which was the more important. But from this result McCracken stated drug activity increases with percentage PSA and that any new drug could be prescreened by this alone.

PSA remains to this day a popular descriptor for use in QSAR. It has been widely accepted by pharmaceutical and medicinal chemists and hence has been used in prediction and modelling of many biological processes, most commonly intestinal absorbance. Much of the pioneering work in PSA in particular with reference to biological absorption has originated from Palm's group at Uppsala University.[92]

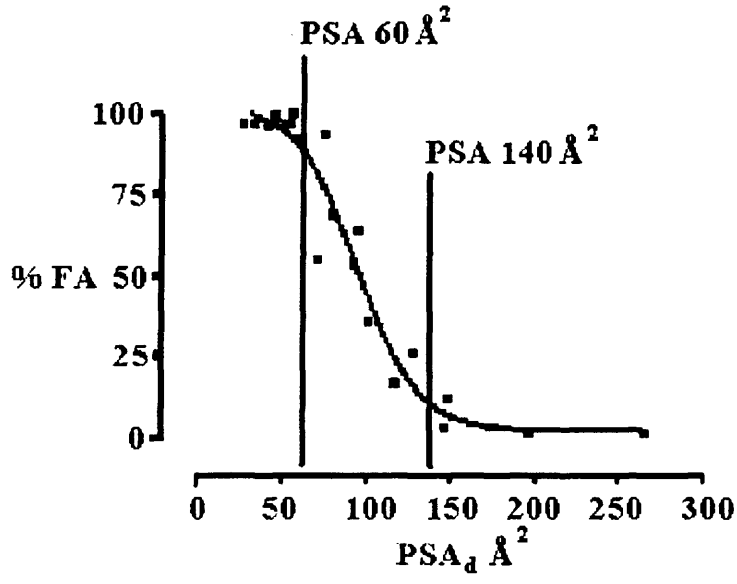
PSA is defined as the surface area of a molecule that arises from N, O, N—H, and O—H atoms, and is simply calculated from a 3D molecular structure usually obtained with some form of energy minimization (see chapter 3.2.2.1).

PSA has been seen to be an excellent descriptor in the modelling of absorption possesses. Using PSA and molecular weight (MW) as descriptors van der Waterbeemd[93] derived a QSAR for passage of 17 molecules across a Caco-2-monolayer.

$$LogP_{app} = 0.008MW - 0.043PSA - 5.165 \quad (1.18)$$

The QSAR showed a good correlation of $R^2 = 0.694$. A further study into absorption was that of Palm,[94] where PSA was used to model fraction of a drug that was absorbed by the intestine (%FA). Using a dataset of 20 molecules that were selected on the basis of being absorbed exclusively through passive diffusion, Palm showed a strong sigmoidal relationship ($R^2 = 0.94$) between PSA_d and %FA. It was seen that when $PSA_d < 60$ Å^2 a molecule would be well absorbed (%FA>90%). While if PSA_d exceeds 140 Å^2 the absorbance was seen as being poor (%FA <10%). The term PSA_d is a measure of dynamic PSA calculated as the Boltzman average of number of conformations (this is discussed in further detail in 1.4.3). This correlation along with the cut-offs is shown in figure 1.1.

Figure 1.1: The Sigmodial fit of PSA_d *vs* %FA.



A number of other studies have also shown PSA to be a suitable descriptor for modelling intestinal absorption, and only in the study of Goodwin[95] were poor correlations seen between PSA and absorption (Caco-2 cells) for a set of 21 peptides.

1.4.2 PSA and Surface Area Type

PSA has been calculated from both the vdw surface area and solvent accessible surface area. Where SASA is applied the algorithm of Lee and Richards[90] is most commonly used to calculate descriptors, formally this corresponds to adding the radius of the solvent molecule to the vdw radii of the atom; essentially a van der Waals surface with inflated radii. Another form of surface that is available is the contact surface, defined as the smooth convex surface traced by the inward facing part of the probe sphere as it rolls over the molecule.[96]

There has been a great deal of debate as to which is the most suitable surface area from which to define PSA. In a study of six beta-blocking agents, Palm[92] showed that PSA_d gives very similar correlations for Caco-2 permeability irrespective of whether obtained from SASA or vdw ($R^2 = 0.99$ and 0.96 respectively). In contrast to this Krarups[97] studies of caco-2 permeation for six beta blocking agents and five prodrugs showed substantially poorer correlations for PSA_d where obtained from vdw surface area ($R^2 = 0.72$) compared to SASA ($R^2= 0.98$). Moreover, another study by Palm[92] of absorption through the rat illium the correlation between $logP_{app}$ in the colonic tissue and the vdw PSA ($R^2=0.91$) was higher than of SASA PSA ($R^2=0.88$)

Water accessible surface area was originally reported to be linearly related to the vdw surface area by Amidion,[88] Palm argued this was only true for small molecules with few intramolecular interactions, and did not hold for all molecules. A good correlation between the two surface areas ($r^2 > 0.83$) was seen for atenolol, but not for alprenolol ($r^2 = 0.45$). Palm also noted that the SASA PSA_d descriptors were larger than the vdw PSA_d for most compounds.

In related QSPR studies where surface area descriptors other than PSA have been applied, similar results have been seen between descriptors derived from SASA and those obtained from vdw surface. In the charged partial surface area QSPR of Stanton[98] for gas chromatographic retention of 107 pyrazines on carbowax 20M, where the probe radius was zero and hence the surface area was effectively the vdw surface, the models were seen to give a slightly poorer fit of R= 0.984 compared to R= 0.994 along with a small

increase in Sd. Stanton concluded that SASA represents the best approximation of the contact surface involved in such interaction.

Amidion[88] stated that on a natural log scale the intercept for such relations, as solubility and molecular surface area should be zero, although this was not reported in his work with SASA. Amidion attributed this to the solvent radii effect noted by Reynolds,[99] which states that when using a solvent probe with a radii of 1.5 Å a theoretical solute molecule with zero radius (no particle) would still give a surface area of 28.3 $Å^2$ which would not give the appropriate intercept. Hence a molecular surface determine with a solvent radius of zero was a more appropriate choice.

1.4.3 Dynamic PSA

The term PSA_d refers to "dynamic PSA", as calculated from a number of conformations, generated through a conformational search. The PSA of each low energy conformation is weighted according to the probability P_J of each conformation J, as given by the normalised Boltzman distribution.

$$P_j = \exp(-\Delta E_j/RT)/\Sigma\exp(-\Delta E_j/RT) \qquad (1.19)$$

Where ΔE_j is the relative steric energy of the J'th conformation, R is the gas constant; T is the temperature in Kelvin.

It has been recommended that for typical drugs 1000 conformations are required to cover all minima and ensure that entire conformational space is explored.[97] The strength of PSA_d is that for flexible molecules where a wide variety of conformations can exist, the effects of these conformations upon surface area are reflected within PSA_d. Palm stated that a method based purely on only the conformation of lowest energy would be sensitive to the choice of force field and hazardous as it may represent only a small part of the conformational space. PSA_d values have been proposed based on vdw surface area by Palm[92,100] and SASA by Krarup.[97]

In order to attempt to add further realism to PSA_d, Palm performed conformational studies on 8 beta blocking agents in simulated water and chloroform environments.[100] Fewer conformations were generated in the simulated chloroform environment and only small differences were seen in the conformations generated in simulated chloroform and vacuum. However low energy conformations generated in simulated water showed a greater degree of intramolecular hydrogen bonding.[101] In general the PSA_d was seen to increase in the order vacuum<chloroform<water as a result of the changes in intramolecular bonding patterns.

Models of caco-2-monolayer permeability for the 8 molecules using PSA_d calculated from each simulated environment were not quantitatively different, with the effect of the simulated solvent acting to only slightly shift the curves of fitting along the PSA_d axis. The use of these simulated environments also dramatically increased the computer time required to generate PSA_d.

A number of authors have suggested that PSA_d may not offer a substantial gain over PSA calculated from a single conformer (static PSA). Clark[101] took the 20 compounds used in the dynamic study of %FA of Palm and calculated static PSA. For molecules with more flexibility there was generally a greater difference between PSA and PSA_d, but the static PSA descriptors gave a correlation with %FA of almost equal value to that of PSA_d. Also, both Palm and Krarup have noted that excellent correlations with experimental data can be obtained from PSA calculated from a single conformation.

Kelder[102] also argued that static PSA obtained from a well built 3D structure will give a PSA of close to that of PSA_d, the exceptions being where hydrophobic collapse or strong intramolecular interactions occur. Stenberg[103] further agreed stating that the PSA of the global minimum conformation generally only differed from PSA_d by a few $Å^2$. In his study of intestinal membrane permeability, PSA and PSA_d were seen to correlate against $logP_{app}$ with values of R^2 of 0.82 and 0.87 respectively. If dynamic PSA is unnecessary then the time taken to calculate PSA is sufficiently small to allow PSA to be used as a pre-screening tool.[81]

1.4.4 Rapid Calculation of PSA

The early success of PSA in modelling biological properties, combined with sheer size of libraries that are produced in combinatorial methods, has led to a number of methodologies for the rapid calculation of PSA.

Clark showed that for the %FA dataset of 20 molecules, an excellent correlation of R^2 of 0.94 could be obtained from 3D structure obtained directly from the program CONCORD[104] with no form of energy minimisation.[1,105] The removal of the energy minimisation (the longest step) from the calculation of PSA meant that values could be generated at a rate of 10+ molecules per second.

A different method to generate PSA as quickly as possible without the need for lengthy geometry optimisation was proposed by Ertl.[106] In this method a type of PSA called topological PSA (TPSA) was defined. TPSA is calculated from a simple summing of a series of tabulated surface contributions of PSA. The method contains a set of 43 polar atom types each with an associated PSA value. The TPSA descriptors were compared to 3D descriptors obtained from CORINA,[107] a simple rule and data based program for generating 3D structure. An R^2 value of 0.982 was reported between the two sets of PSA, the majority of outliers were large molecules with many polar atoms.

The TPSA descriptors were seen to be equally successful in correlation of biological properties when used to remodel six published datasets. Even more surprising, the descriptor were seen to give comparable R^2 values to PSA_d descriptors. (model of Caco-2 absorbance for 20 molecules $R^2 = 0.91$ cf. $R^2 = 0.94$) The un-computationally demanding nature of this method makes it extremely rapid with 8000 TPSA values produced per minute. Egan and Lauri[108] and Labatue[109] have proposed two other rapid PSA calculation methods from summation of fragments.

1.4.5 PSA and Additional Descriptors

The use of PSA as a descriptor reduces the various ways a molecule can interact with the environment to a single number. Hansch[20] established that a single descriptor may not be sufficient to explain a biological or physiochemical property as it may be reliant on more that one interaction. Work by Abraham and others[79,81,110] demonstrate that the relative importance of these interactions can differ greatly. Hence it can be seen that a single number e.g. PSA or $logP_{oct}$ cannot hope to model all such properties.

In a model of blood brain barrier penetration for 57 molecules, Clark[111] noted that a single parameter model of PSA was incapable of predicting the varying penetrative strengths of nonpolar molecules. Clark attempted to construct models incorporating molecular weight, molecular volume, nonpolar surface area and ClogP: only ClogP was seen to significantly improve the model.

Non-polar surface area has been very successful in combination with PSA for modelling intestinal absorption, and both Krarup[97] and Palm[94] included non-polar surface area in models of intestinal absorption. The contributions from non-polar surface area were investigated by Stenburg for a series of 19 oligopeptides.[112] A sigmodial fit was seen for a model of PSA and non-polar surface area, and an excellent model was produced which out-performed similar equations that were reliant on experimentally observed descriptors, such as octanol-water and heptane-water partition.

Winiwarter[113] attempted numerous models of the human effective intestinal permeability for 13 molecules using PSA and a number of additional descriptors. A one parameter model of just PSA gave an R^2 value of 0.76. On addition of a second descriptor, the number of hydrogen bond donors, the R^2 value increased to 0.88. Further improvements were made to the model by inclusion of either logP or ClogP, giving R^2 values of 0.98 for both descriptors. Other descriptors such as HOMO and LUMO energies, molecular weight, dipole moment and total number of atoms were seen to be less significant upon intestinal permeability when used in conjunction with PSA.

Stenberg *et al*[114] fragmented the PSA and molecular surface area in to separate surface area descriptors which included the surface area of hydrogen attached to oxygen, sp3 carbon, sp3 nitrogen, saturated non polar and double bonded oxygen. These descriptors were used to create a model of intestinal absorbance. This model, which they called the partitioned total surface area (PTSA), was a marked improvement over traditional PSA methods giving results comparable to those obtained from more computationally demanding methods such as quantum mechanical calculations.

In a review of prediction of physiochemical properties of dug like molecules Blake[115] noted that no distinction has been made in PSA between the hydrogen bond acceptor or donor contributions of PSA, an area that Blake noted that could warrant further studies.

1.4.6 What Does PSA Represent?

PSA has met a great deal of success in modelling passive absorbance, it has been shown from previous QSAR studies that the two key components of passive absorbance are lipophilicity and hydrogen bonding potential[116], so PSA must be a measure of at least one of these factors. If PSA of a homologous series of acids is considered, lipophilicity ($logP_{oct}$) increases with increasing chain length while PSA remains constant, so logically PSA is a measure of hydrogen bonding. This can be further rationalised when one considers the number of models in which logP has been used in combination with PSA to predict passive absorbance.

Palm[100] provided further evidence that PSA was a measure of hydrogen bonding by showing the high correlation between PSA_d and the number of hydrogen bonds that could be formed by a molecule (R=0.92). Palm also pointed out that PSA is a more informative measure of hydrogen bonding than many other theoretical approaches of calculating hydrogen bonding as the influence of the 3D structure can account for such effects as shielding and burial of polar atoms in a molecule.

Stenberg[117] proposed a deconvolution of PSA to facilitate the interpretation of the composite descriptor, and to suggest methods for faster calculation or more easily obtainable substitutes. He calculated a number of hydrogen bond donor and acceptor

properties using the program HYBOT[118] and MOLSURF[119] and correlated them with PSA using PLS for a set of 128 molecules. The program HYBOT created by Raevsky and co-workers uses a large database of thermodynamic data relating to hydrogen bonding to calculates free energy hydrogen bond donor (ΣC_b) and acceptor strengths (ΣC_a) for a given molecule;[120] numerous absorption and permeability datasets have been modelled with HYBOT descriptors.[121] The program MOLSURF uses the wavefunction to compute various properties related to the molecular valence region. MolSurf descriptors describe properties such as hydrogen bonding, polarity and polarisability. It should be noted that MOLSURF is too computationally demanding to be applied to large libraries or other scenarios in which rapid calculation of descriptors are required.

Stenburg[117] revealed a good correlation for hydrogen bond acceptor strength and PSA, while poor correlations were seen with hydrogen bond donor strength. Polarity and size related properties were seen to be of less importance. Most importantly, Stenburg showed the number of hydrogen bonds and not their strength was most important to PSA. The highest correlation with PSA were seen with descriptors for the number of H-bonding acceptor oxygen atoms, number of H-bond acceptor nitrogen atoms and the total number of hydrogen bond donors, which described 93% of the variance of PSA. Ostberg and Norinder[122] also found that simple counts of acceptor nitrogen and oxygen atoms plus the sum of hydrogen atoms bound to N and O correlate strongly with PSA > $R^2 = 0.93$.

These conclusions were drawn from single conformations, and Stenburg noted the use of PSA_d in this analysis was not possible, as the computations would be too demanding. Hence the effects of intramolecular hydrogen bonds, which would represent a substantial problem to prediction of PSA by atom or fragment counts, may be less pronounced. Stenberg also noted that simple atom and fragment counts might not be extensive enough to define PSA for large flexible molecules with many polar atoms.

1.4.7 PSA and Hydrogen Bonding Strength

The evidence indicates that PSA is a representation of hydrogen bonding but not hydrogen bonding strength. Possibly the most obvious reason for this is the grouping together of all polar atoms N, O, N—H, and O—H as having the same contribution to

PSA. It is well known[123] that the H-bond strength is far from uniform for different functional groups. For example donor ability of N—H atoms can differ by an order of magnitude (*e.g.* for dimethylamine A = 0.08, tetrazole A = 0.88 using Abraham's scale of hydrogen bond acidity), and taking these groups' contribution to PSA as identical must introduce errors.

The inability of PSA to account for the varying hydrogen bonding strengths has been noted by a number of authors. In the study of Krarup[97] the two nitrogens in the 1,2,5 thiadiazole ring of timodol and pro timidol were not included in their definition of PSA as they were seen to be reluctant hydrogen bond acceptors.[101] This non-contribution of specific polar atoms was also discussed by Clark,[101] where he pointed out that crystal structure surveys and *ab initio* calculations[124-126] show the 'ether' oxygen in a ester is only rarely a hydrogen bond acceptor, and should perhaps also be omitted from the PSA calculation altogether.

A number of other authors have commented on the possible unrealistic assumptions made when calculating PSA and treating it as a measure of hydrogen bonding. Both Ertl [106] and Blake [115] commented that a more realistic approach to calculating PSA would be to scale these to account for hydrogen bonding strength.

1.4.8 Charged Partial Surface Areas

One method that has been suggested to increase the accuracy and realism of molecular surface area descriptors has been through the use of charged partial surface areas, In this method, the surface area (either vdw or SASA) is calculated and then scaled using a value derived from the electrostatic charge. This has been done in numerous studies although to our knowledge never directly related to PSA.

Stanton[98] proposed a number of charge partial surface area descriptors, which were grouped by charge type thus, partial positive surface area descriptors (PPSA) and partial negative descriptors (PNSA). Descriptors to describe differences in charge (DPSA) functionally charged descriptors (FPSA) and a similar set of total surface weighted partial surface area descriptors (WPSA and WNSA) were also proposed. Two descriptors were

also included to describe the most highly charged negative and positive atom (RPCG and RNCG).

Using Stepwise multiple linear regression with other descriptors such as number of single bonds, molecular polarisability for gas chromatographic retention of 107 pyrazines on Carbowax 20M, a six parameter QSAR was defined which had a R^2 value of 0.988 and a Sd of 32.9. In this equation three of the descriptors were charged partial surface area descriptors.

1.5 Conclusions

Numerous models and descriptors have been developed to predict and describe biological and physiochemical properties. The use of 3D QSAR has proven an area of particular interest as features such as steric hindrance and burial of important molecular features can be explained. The use of molecular surfaces has met with a great deal of success as a 3D modelling tool, and one such molecular surface descriptor (PSA) has show great potential for modelling biological possesses.

Numerous models of PSA have been proposed along with different routes to its calculation i.e. type of surface from which it is defined. A dynamic form of PSA has also been proposed in which the effects of multiple low energy conformations are considered. There is a substantial disagreement as to which form of PSA is the most accurate, although the general consensus is that dynamic PSA offers only marginal improvement over PSA calculated from a single conformation. One area in which all authors are in agreement is that the time consuming nature of PSA_d makes it an unsuitable descriptor for virtual screening of large libraries.

While PSA has met with a great deal of success it is not without its failings. Most notably, it does not accurately represent hydrogen bonding as both donors and acceptors are classed together and the relative hydrogen bond strengths of different functional groups is ignored. PSA as a descriptor is also fairly uninformative compared to methods such as LFER and interpretation of the QSAR produced tend to provide little information.

1.6 References

1. S.D. Pickett, I.M. Mclay, D.E. Clark, *J. Chem. Inf. Comput. Sci.*, **2000**. 40. 263.
2. M.T.D. Cronin, J.C. Dearden, G.P. Moss, G. Murray-dickson, *Eur. J. Pharm. Sci.*, **1999**. 7. 325.
3. A. Crum-Brown, T.R. Frasier, *Trans.Roy.Soc.Edinburgh*, **1898**. 25. 151.
4. B.J. Richardson, *Med.Times and Gazette*, **1869**. 2. 703.
5. E. Overton, *Studien uder die Narkose, Fisher, Jena, Germany*, **1901**.
6. H. Meyer, *Arch.Exp.Pathol.Pharmakol*, **1901**. 46. 338.
7. K.H. Meyer, H. Gottlieb-Billroth, *Z. Physiol. Chem.*, **1935**. 112. 55.
8. R. Collander, H. Barlund, *Ebenda*, **1933**. 11. 82.
9. R. Collander, *Acta Chem. Scand*, **1947**. 13. 363.
10. R. Collander, *Acta. Chem. Scand*, **1951**. 8. 774.
11. A. Leo, c. Hansch, *J. Org. Chem.*, **1971**. 36. 1539.
12. L.P. Hammet, *Chem. Rev*, **1935**. 17. 125.
13. L.P. Hammett, *Trans. Faraday Soc.*, **1938**. 34. 156.
14. L.P. Hammett, *Physical Organic Chemistry*. 1940, New York: McGraw-Hill.
15. J. Shorter, *Pure Applied Chem.*, **1994**. 66. 2451.
16. M. Charlton, *Progr. Phys. Org. Chem.*, **1981**. 13. 119.
17. C. Hansch, A. Leo, R.W. Taft, *Chem. Rev.*, **1991**. 91. 165.
18. R.W. Taft, *Steric effects in organic chemistry*, ed. M.S. Newman. 1956, New York: Wiley.
19. J.D. Roberts, W.T. Moreland, *J. Am. Chem. Soc.*, **1953**. 75. 2167.
20. C. Hansch, *Acc. Chem. Res.*, **1969**. 2. 232.
21. C. Hansch, *Acc. Chem. Res.*, **1993**. 26. 147.
22. R.D. Cramer, D.E. Patterson, J.D. Bruce, *J. Am. Chem. Soc.*, **1988**. 110. 5959.
23. R.D. Cramer, *Abstracts of Papers of the American Chemical Society*, **2002**. 224. 056.
24. R.D. Cramer, *Journal of Medicinal Chemistry*, **2003**. 46. 374.
25. L.E. Kiss, I. Kovesdi, J. Rabai, *Journal of Fluorine Chemistry*, **2001**. 108. 95.
26. M.D. Wessel, P.C. Jurs, J.W. Tolan, S.M. Muskal, *J. Chem. Inf. Comput. Sci.*, **1998**. 38. 726.
27. P.D.T. Huibers, V.S. Lobanov, A.R. Katritzky, D.O. Shah, M. Karelson, *Langmiur*, **1993**. 12. 1462.
28. P.D.T. Huibers, V.S. Lobanov, A.R. Katritzky, D.O. Shah, M. Karelson, *J. Colloid Interface Sci,*, **1997**. 187. 113.
29. A.R. Katritzky, V.S. Lobanov, M. Karelson, *Chem. Soc. Rev.*, **1995**. 24. 279.
30. B.E. Mitchell, P.C. Jurs, *J. Chem. Inf. Comput. Sci.*, **1998**. 38. 489.
31. R.F. Rekker, *The Hydrophobic fragmental Constant*. 1977, New York: Elsevier.
32. C. Hansch, A. Leo, D. Hoekman, *Exploring QSAR Hydrophobic Electronic, and steric Constants*. 1995, Washington DC: American Chemical Society.
33. C. Hansch, J. Dunn, *J. Pharm. Sci*, **1972**. 61. 1.
34. H. Kubinyi, *Progr. Drug. Res.*, **1979**. 23. 97.
35. C. Hansch, J.M. Clayton, *J. Pharm. Sci*, **1973**. 72. 1266.
36. S.S. Davis, T. Higuchi, J.H. Rytting, *Adv. Pharm. Sci*, **1974**. 4. 73.
37. J.K. Seydel, K.J. Schaper, *Pharmacol. Ther.*, **1982**. 15. 131.
38. S.H. Yalkowsky, S. Banerjee, *Aqueous Solubility. Methods of Estimation for Organic Compounds*. 1992, New York: Dekker.

39. C. Hansch, R.M. Muir, T. Fujita, P.P. Maloney, F. Griger, F. Streich, *J. Am. Chem. Soc.*, **1963**. 85. 2817.
40. P. Buchwald, N. Bodor, *Current. Med. Chem.*, **1998**. 5. 353.
41. C. Hansch, A.J. Leo, *Substituent Constants for Correlation Analysis in Chemistry and biology*. 1979, New York: Wiley.
42. J. Sangster, *J. Phys. Chem. Ref. Data.*, **1989**. 18. 1111.
43. N.P. Franks, W.R. Lieb, *Nature*, **1978**. 274. 339.
44. W.D. Stein, *Transport and Diffusion across Cell Membranes*. 1986, Orlando: Academic Press.
45. R. Mannhold, K. Dross, *Quantitative Structure-Activity Relationships*, **1996**. 15. 403.
46. H. vandeWaterbeemd, R. Mannhold, *Quantitative Structure-Activity Relationships*, **1996**. 15. 410.
47. R. Mannhold, R.F. Rekker, C. Sonntag, A.M. Terlaak, K. Dross, E.E. Polymeropoulos, *Journal of Pharmaceutical Sciences*, **1995**. 84. 1410.
48. A.J. Leo, *Chemical Reviews*, **1993**. 93. 1281.
49. H. van-de-Waterbeemd, *Hydrophobicity of Organioc Compounds: how to calculate it by Personal Computers*. 1986, Vienna: CompuDrug.
50. J. Sangster, *Octanol Water Partition Coefficents: Fundamentals and Physical Chemistry*. 1997, Chichester: Wiley and Sons.
51. G.G. Nys, R.F. Rekker, *Chim. Ther.*, **1973**. 8. 521.
52. G.G. Nys, R.F. Rekker, *Chim. Ther.*, **1974**. 9. 361.
53. T. Suzuki, Y. Kudo, *J. Comput-Aided Mol. Des*, **1990**. 4. 155.
54. A. Leo, P.Y.C. Jow, C. Silipo, *J. Med. Chem.*, **1975**. 18. 865.
55. W.J. Lyman, *Handbook of Chemical Property Estimation Methods*. 1982, New-York: McGraw-Hill.
56. A. Leo, P.Y.C. Jow, C. Silipo, *J.Chem. Soc. Perkin Trans 2.*, **1983**. 825.
57. R. Calvino, A. Gasco, A.J. Leo, *J. Chem. Soc. Perkin Trans. 2*, **1992**. 1643.
58. P.J. Taylor, *Comprehensive Medicinal Chemistry*. Vol. 4. 1990, New York: Pergamon Press. 241.
59. A.K. Ghose, G.M. Crippen, *J. Comput. Chem.*, **1986**. 7. 565.
60. A.K. Ghose, A. Pritchett, G.M. Crippen, *J. Comput. Chem.*, **1988**. 9. 80.
61. V.N. Viswanadhan, M.R. Reddy, R.J. Bacquet, M.D.J. Erion, *J. Comput. Chem.*, **1993**. 14. 1019.
62. P. Broto, G. Moreau, C. vandycke, *Eur. J. Med. Chem.*, **1984**. 19. 71.
63. M. Karelson, V.S. Lobanov, A.R. Katritzky, *Chem. Rev.*, **1996**. 96. 1027.
64. G. Klopman, L.J. Iroff, *J. Comput. Chem.*, **1981**. 2. 157.
65. N. Bodor, M.J. Huang, *Journal of Pharmaceutical Sciences*, **1992**. 81. 272.
66. N. Bodor, Z. Gabanyi, C.K. Wong, *Journal of the American Chemical Society*, **1989**. 111. 3783.
67. Y. Sasaki, H. Kubodera, T. Matuszaki, H. Umeyama, *J. Pharmacobio. Dyn.*, **1991**. 14. 207.
68. G. Bravi, J.H. Wikel, *Quant. Stuct. Act. Relat.*, **2000**. 19. 39.
69. K.H. Kin, *J. Comp. Aid. Des.*, **1995**. 9. 308.
70. M.J. Kamlet, R.W. Taft, *J. Am. Chem. Soc.*, **1976**. 98. 377.
71. R.W. Taft, M.J. Kamlet, *J. Am. Chem. Soc.*, **1976**. 98. 377.
72. R.W. Taft, M.H. Abraham, R.M. Doherty, M.J. Kamlet, *Nature*, **1985**. 313. 384.
73. M.J. Kamlet, L.M. Abboud, R.W. Taft, *Progr. Phys. Org. Chem.*, **1981**. 13. 485.
74. M.J. Kamlet, J. Abboud, M.H. Abraham, *J. Org. Chem.*, **1983**. 48. 2877.

75. R.W. Taft, J.L.M. Abboud, M.J. Kamlet, M.H. Abraham, *J. Soln. Chem.*, **1985**. 14. 153.
76. M.J. Kamlet, R.M. Doherty, J.L.M. Abboud, M.H. Abraham, R.W. Taft, *CHEMTECH*, **1986**. 566.
77. M.J. Kamlet, M.H. Abraham, R.M. Doherty, R.W. Taft, *J. Am. Chem. Soc.*, **1984**. 106. 464.
78. R.W. Taft, M.H. Abraham, G.R. Famini, R.M. Doherty, J.L.M. Abboud, M.J. Kamlet, *J. Pharm. Sci*, **1985**. 74. 807.
79. M.H. Abraham, *Chem. Soc. Revs.*, **1993**. 22. 73.
80. M.H. Abraham, J.C. McGowan, *Chromatographia*, **1987**. 23. 243.
81. J.A. Platts, M.H. Abraham, *Environ. Sci. Technol.*, **2000**. 34. 318.
82. J.A. Platts, D. Butina, M.H. Abraham, A. Hersey, *Journal of Chemical Information and Computer Sciences*, **1999**. 39. 835.
83. M.H. Abraham, H.S. Chadha, G.S. Whiting, R.C. Mitchell, *J. Pharm. Sci.*, **1994**. 83. 1085.
84. M.J. Kamlet, R.M. Doherty, M.H. Abraham, Y. Marcus, R.W. Taft, *J. Phys. Chem.*, **1988**. 92. 5244.
85. N. El-Taylor, R.S. Tsai, B. Testa, P.A. carrupt, A. Leo, *J. Pharm. Sci*, **1991**. 80. 590.
86. I. Langmuir, *Third Colloid Symposium Monograph*. 1925, New York: Chemical Catalog Co. 3.
87. R.B. Hermann, *J. Phys. Chem.*, **1971**. 76. 2754.
88. G.L. Amidion, S.H. Yalkowsky, S.T. Anik, S.C. valvani, *J. Phys. Chem.*, **1975**. 79. 2239.
89. W.J. Dunn, M.G. Koehler, S. Grigoras, *J. Med. Chem*, **1987**. 30. 1121.
90. B. Lee, F.M. Richards, *J. Mol. Biol*, **1971**. 55. 379.
91. R.O. McCracken, K.B. Lipkowitz, *J. Parasitol*, **1990**. 76. 853.
92. K. Palm, K. Luthman, A.L. Ungell, G. Strandlund, P. Artursson, *Journal of Pharmaceutical Sciences*, **1996**. 85. 32.
93. H. vandeWaterbeemd, G. Camenisch, G. Folkers, O.A. Raevsky, *Quantitative Structure-Activity Relationships*, **1996**. 15. 480.
94. K. Palm, P. Stenberg, K. Luthman, P. Artursson, *Pharmaceutical Research*, **1997**. 14. 568.
95. J.T. Goodwin, B. Mao, T.J. Vidmar, R.A. Conradi, P.S. Burton, *J. Pept. Res*, **1999**. 56. 355.
96. F.M. Richards, *Annu. Rev. Biophys. Bioeng.*, **1977**. 6. 151.
97. L.H. Krarup, I.T. Christensen, L. Hovgaard, S. Frokjaer, *Pharmaceutical Research*, **1998**. 15. 972.
98. D.T. Stanton, P.C. Jurs, *Anal. Chem.*, **1990**. 62. 2323.
99. J.A. Reynolds, D.B. Gilbert, C. Tanford, *Proc. Natl. Acad. Sci. U.S.A.*, **1974**. 71. 1974.
100. K. Palm, K. Luthman, A.L. Ungell, G. Strandlund, F. Beigi, P. Lundahl, P. Artursson, *Journal of Medicinal Chemistry*, **1998**. 41. 5382.
101. D.E. Clark, *Journal of Pharmaceutical Sciences*, **1999**. 88. 807.
102. J. Kelder, P.D.J. Grootenhuis, D.M. Bayada, L.P.C. Delbressine, J. Ploenmen, *Pharm. Res.*, **1999**. 16. 1514.
103. P. Stenberg, K. Luthman, P. Artursson, *Pharmaceutical Research*, **1999**. 16. 205.
104. R.S. Pearlman, *Chem. Des. Auto.*, **1987**. 1. 1.
105. D.E. Clark, *Comb. Chem. High Throughput Screen.*, **2001**. 4. 447.
106. P. Ertl, B. Rohde, P. Selzer, *Journal of Medicinal Chemistry*, **2000**. 43. 3714.

107. J. Gasteiger, J. Sadowski, J. Schuur, P. Selzer, L. Steinhauer, V. Steinhauer, *Journal of Chemical Information and Computer Sciences*, **1996**. 36. 1030.
108. A. Cheng, D.J. Diller, S.L. Dixon, W.J. Egan, G. Lauri, K.M. Merz, *J. Comput. Chem.*, **2002**. 23. 172.
109. P. Labute, *J. Mol. Graph.*, **2000**. 18. 464.
110. H.v.d. Waterbeemd, M. Kansy, *Chimia*, **1992**. 46. 299.
111. D.E. Clark, *Journal of Pharmaceutical Sciences*, **1999**. 88. 815.
112. P. Stenburg, K. Luthman, H. Ellens, C.P. Lee, P.L. Smith, A. Lago, J.D. Elliott, P. Artursson, *Pharm. Res.*, **1999**. 10. 1520.
113. S. Winiwarter, N.M. Bonham, F. Ax, A. Hallberg, H. Lennernas, A. Karlen, *J. Med. Chem*, **1998**. 41. 4939.
114. P. Stenberg, U. Norinder, K. Luthman, P. Artursson, *Journal of Medicinal Chemistry*, **2001**. 44. 1927.
115. J.F. Blake, *Curr. Opin. Biotechnol.*, **2000**. 11. 104.
116. R.A. Conradi, P.S. Burton, R.T. Borchardt, *Lipophilicity in drug action and toxicology*. 1996: Weinhiem.
117. P. Stenberg, U. Norinder, K. Luthman, P. Artursson, *J. Med. Chem*, **2001**. 44. 1927.
118. HYBOT, *pION 5 Constitution Way, Woburn, MA*.
119. MOLSURF, *Qemist AB, Hertig Carls Alle 29, Sweden Clark III*.
120. O.A. Raevsky, V.Y. Grigor'ev, D.B. Kireev, N.S. Zefirov, *Quant. Struct. Act. Relat*, **1992**. 11. 49.
121. O.A. Raevsky, K.J. Schaper, *Eur. J. Med. Chem.*, **1998**. 33. 799.
122. T. Osterberg, U. Norinder, *J. Chem. Inf. Comput. Sci*, **2000**. 16. 205.
123. M.H. Abraham, M. Berthelot, C. Laurence, P.J. Taylor, *J. Chem. Soc., Perkin Trans., 2*, **1998**. 1. 187.
124. H.J. Bohm, S. Brode, U. Hesse, G. Klebe, *Chem.-Eur. J.*, **1996**. 2. 1509.
125. I.J. Bruno, J.C. Cole, J.P.M. Lommerse, R.S. Rowland, R. Taylor, M.L. Verdonk, *Journal of Computer-Aided Molecular Design*, **1997**. 11. 525.
126. P.R. Rablen, J.W. Lockman, W.L. Jorgensen, *Journal of Physical Chemistry A*, **1998**. 102. 3782.

Chapter 2. Statistical Methods

A number of different statistical methods have been used in QSAR and QSPR studies, the following chapter outlines some of the more commonly used statistics and those used in the subsequent studies. Further information on all of these techniques can found in the following references.[1-7]

2.1 Linear Regression

One of the most common methods of deriving QSAR equations is via linear regression; this method uses least squares fitting to find the best combination of coefficients within the QSAR equation. To demonstrate the mechanism of the least squares we shall first consider the simplest case where the property we wish to model is the function of just one descriptor. In this case the equation that we need to define is.

$$Y=MX+C+E \qquad (2.1)$$

Where Y is the observed property we wish to model, e.g. logP, referred to as the dependant variable. X is the descriptor, e.g. molecular volume, and is referred to as the independent variable. M and C are the coefficients, and E is a random error term, which is removed when the equation is used predictively. The goal of regression analysis is to calculate the optimum value of M and C that will minimise the sum of deviation of the observations from the fitted equation. Finding the equation that gives the relationship between X and Y is known as finding the regression line. The values of X and Y occur in pairs. For n values of X we would have values X_1, X_2 ..... X_n which correspond to the Y values Y_1, Y_2 .... Y_n. The line of best fit for these n corresponding points is that which minimises the sum of squares deviations of the predicted Y values from the observed Y values. The equation of the best fitting line is thus

$$\hat{Y} = MX + (\overline{Y} - M\overline{X}) \qquad (2.2)$$

Where

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \qquad (2.3) \qquad \overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} \qquad (2.4)$$

And

$$X = \frac{\sum_{i=1}^{n} (X_i - \overline{X})Y_i}{\sum_{i-1}^{n} (X_i - \overline{X})^2} \qquad (2.5)$$

The coefficient M is the slope of the regression line and $(\hat{Y} - M\overline{X})$ is the Y intercept. The symbol $\hat{Y}$ indicates that this is the value Y as predicted by the equation. Once the equation of the line of best fitting is determined a value of $\hat{Y}_i$ can be predicted by inserting the appropriate X_i value in to the equation

$$\hat{Y}_i = MX_i + (\overline{Y} - M\overline{X}) \qquad (2.6)$$

The most efficient formula for the calculation of M is

$$M = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}} \qquad (2.7)$$

The quality of a linear regression equation is most often reported as the squared correlation coefficient, or r^2 value. This coefficient is the fraction of total variation in the dependant variables that is explained by the regression equation. To determine r^2 it is necessary to calculate the total sum of squares *(TSS)* of the deviation of the observed Y values from the average the mean $\overline{Y}$ and the explained sum of squares *(ESS)*, which is the sum of deviation of the values $\hat{Y}_i$ calculated from the model from the mean $\overline{Y}$. Another term that is commonly used in calculation of r^2 is the residual sum of squares *RSS*, which is the square of the difference between the observed and calculated values of Y. the

difference in value between the observed and calculated values of Y is referred to as the residual it is the measure of how accurately a value of $\hat{Y}$ is predicted, The TSS is the sum of RSS and ESS. r^2 is calculated from these sum of squares values thus.

$$r^2 = \frac{ESS}{TSS} \equiv \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} \qquad (2.8)$$

Values of r^2 range from 0 to 1, a value of one indicates that all of the variance of the observed data is being modelled by variation in the independent variable and a value of 0 indicates none of the variance has been explained.

2.2 Multiple Linear Regression

In multiple linear regression, the techniques of linear regression are expanded so that the effects of more than one independent variable can be modelled simultaneously. Where as linear regression fits the value of a line in two-dimensional space, multiple linear regression fits a multidimensional surface. For a series of p independent X variables the general form of the multiple linear regression is as follows

$$Y = C + M_1X_1 + M_2X_2 \ldots M_pX_p + E \qquad (2.9)$$

Similarly to equation 1 C is the Y intercept, $M_1 \ldots M_P$ are the gradients of the descriptors $X_1 \ldots X_P$. E is a random error term, when the equation is used predictively this term is omitted. In order to calculate values for $M_1\ M_2 \ldots M_P$ and C the equation is treated using matrix algebra, which produces the matrix form.

$$\boldsymbol{Y = XM + E} \qquad (2.10)$$

Where ***Y, M***, and ***E*** are vectors, and ***X*** is a matrix (called the model matrix or design matrix). The vectors ***Y M*** and ***E*** and matrix ***X*** for equation 2.10 would be as follows

$$X = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ 1 & X_{21} & X_{22} & \cdots & X_{2p} \\ 1 & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix} \quad Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \quad E = \begin{pmatrix} E_1 \\ E_2 \\ \vdots \\ E_n \end{pmatrix} \quad M = \begin{pmatrix} C \\ M_1 \\ \vdots \\ M_p \end{pmatrix}$$

Values of C and $M_1 \ldots\ldots M_2$ are calculated by least squared analysis, for n observations, the residual is minimised as follows.

$$\sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} \left(Y_i - C - M_1 X_{i1} - M_2 X_{i2} - \ldots M_P X_{ip}\right)^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 \qquad (2.11)$$

The quality of the fit of MLRA is indicated by the multiple correlation co-efficient R^2, this statistic is analogous in calculation and interpretation to r^2.

2.3 Root Mean Square Error

The route mean square error (RMSE) represents the square root of the mean of deviation from the mean. This is used to assess the accuracy of the values calculated by the model.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2}{n}} \qquad (2.12)$$

A similar measure is standard deviation (Sd). This is almost identical to RMSE except the mean of squares is divided by n-1. This is done to prevent the underestimation of the total population variance. Although for datasets where $n > 20$ the difference in the two statistical methods becomes so small that either can be used to calculate the variation in the data.

2.4 F-ratio

Another statistical validation of MLRA is f-ratio. The f-ratio indicates the likelihood that relationship is not one derived by chance. The f-ratio is dependent upon the number of independent variables and the number of data points within the equation. The value which corresponds to a particular level of confidence in the model falls as the number of data points increases and or the number of independent variables falls. The reason for this is that an equation would be expected to have greater predictive power if it is predicting a large number of data points with the fewest possible descriptors. The numbers of degrees of freedom associated with each parameter are used to take this into account. A simple linear regression is associated with n-1 degrees of freedom as the fitted line always passes through the means of the dependant and independent variables. The total sum of squares is associated with N-1 degrees of freedom. If there are P independent variable then there are N –P –1 degrees of freedom associated with the explained sum of squares.

$$F - ratio = \frac{ESS}{P} \frac{N - P - 1}{RSS} \tag{2.13}$$

2.5 Cross-validated R^2 and Training and Test Sets

The cross-validated R^2 (R^2_{cv}) is a measure of the internal self-consistency of a model and reflects the predictive power of the model. The value of R^2 is determined by removing a value from the dataset, deriving a model from the remaining data and then using this model to predict the value of the excluded value. This predicted value is then compared to the observed value. This is repeated until every point in the model has been excluded and its value has been predicted. Values obtained for R^2 are usually higher than those obtained for R^2_{cv}. Another measure of predictive accuracy is the PRESS statistic (predicted residual sum of squares) this value is similar to the residual sum of squares although the values of Y_i calc are obtained from models which do not include the corresponding values of X_i.

$$R_{cv}^2 = 1 - \frac{PRESS}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2} \tag{2.14}$$

$$PRESS = \sum_{i=1}^{N}(\hat{Y}_i - Y_i)^2 \tag{2.15}$$

A further more robust test of the models predictive accuracy is through the use of training and test sets. In this method a large number of data points, usually about 20 – 25% of the total number of data points in the dataset, are removed and called the test set. A model is then created from the remaining points, which are called the training set, and the values of the test set are predicted. This process can be repeated several times until all compounds have been excluded and predicted.

2.6 Interpretation of MLRA Equations

While the ability of QSAR equations to predicted properties is useful, the ability to glean information from the models about the properties governing the modelled system should not be overlooked.

Where all dependant variables are scaled similarly it is possible to gain information about the significance and role of each descriptor directly from the regression coefficient. However, this is not possible where the values of the individual descriptors vary greatly in size, for models where this is the case the significance of individual descriptors can be assessed using the t-ratio. The t-ratio is obtained by dividing the relevant regression coefficient (M) by the standard error of the coefficient s(M).

$$t - ratio = \frac{M}{s(M)} \tag{2.16}$$

The significance of the t-ratios is informative. A large t-ratio value shows that a descriptor is accounting for a significant proportion of the variation. Similarly a small t-ratio value would indicate that the descriptor is insignificant, if the t-ratio is sufficiently small and the descriptor is seen not to be contributing to the model it may be removed from the regression.

Where the descriptors are physically meaningful it is possible to compare the relative size of all t-ratios and determine what factors are most strongly governing the modelled system. Not only is the value of a t-ratio informative but its sign also provides information on the role of the descriptor within the model. For example in a model for the prediction of $\log P_{oct}$ a positive value for the t-ratio of the molecular size descriptor would indicate that large molecules would more preferentially be drawn into the octanol, alternatively if the molecular size descriptor had a negative t-ratio it would tell us that partition of larger molecules into octanol is unfavourable. Where the properties of a system as determined from t-ratio values reflect the known physical properties of the system further validity is added to the model.

It should be stated that assumptions made on t-ratio values should not be made lightly as a limitation of all regression techniques is that one can only ascertain relationships, but never be certain about underlying causal mechanism.

2.7 Stepwise regression

When a large number of descriptors are generated it is often difficult to know which are important within the regression. A solution to this is the use of stepwise regression. There are two methods of stepwise regression forward and backward stepping. In forward stepping an equation is derived with one descriptor, which is seen to make the most contribution, based on its t-ratio. In the next step the next most influential descriptor is added to the equation, this is repeated a number of times until all significant descriptors have been included into the regression. Backward stepping is simply the opposite of this where an equation is calculated with all the descriptors with the least significant descriptor being ejected from the equation at each step.

2.8 Outliers

Where a point is seen to be modelled poorly and large residuals are reported the point is called an outlier. This is occasionally defined rigidly as any point whose residual is three times that of the RMSE, although it is also defined as any point whose residual is

substantially larger than other residuals in the model. In extreme cases outliers can seriously bias the results by "pulling" or "pushing" the regression line in a particular direction, thereby leading to biased regression coefficients. The effects of outliers are particularly influential on the values of R^2_{cv}.

There can be number of different reasons for outliers, most obviously that the experimentally obtained value is incorrect due to mistakes in measurement or calculation. It is occasionally possible to corroborate such mistakes by comparing experimentally observed data for a homologous series of molecules.

Where observational error is not the cause of an outlier the fault logically lies within the model, although this may not mean the model is incorrect merely that the boundaries in which it operates have been violated. For example if a model were created for passive diffusion in a biological system, any molecule with strong active transport properties would be seen as an outlier. Outliers can also be generated where a molecule has a descriptor or property that greatly exceeds the boundaries of the model. Structural abnormalities and properties that are not permitted by the model can also be a large source of outliers e.g. charged molecules and zwitterions.

An outlier should only be removed if there is a justification for its removal. Often excluding just a single extreme case can yield a completely different set of results. When the outlier has no structural abnormalities and reliable experimentally observed values the QSAR should be re evaluated.

2.9 Limitations to MLRA and Alternatives

While MLRA is a powerful tool it is not without its limitations. Fitting problems can occur when the data does not have a good distribution of values. Datasets where the distribution of values is irregular can generate models where the line is being fitted through a few isolated pockets of data. While the model will show good accuracy when analysed with statistical tests such as R^2, it is likely to be predictively poor as the line is not fitting the individual points in each pocket, instead the line of best fit is only

intersecting with the pocket. These pockets are often easily observable in a plot of observed versus calculated data.

MLRA is also limited in the number of descriptors that can be applied to a dataset and still achieve a statistically significant result. It has been stated that at a minimum there must be five data points for every descriptor, and that the optimum ratio to gain a stable model is at least 10 to 20 times as many observations than descriptors.

Another limitation to MLRA is where descriptors are highly correlated with each other ($R^2 > 0.5$). This internal correlation causes redundancy within the descriptors as the same direction of fit is being modelled twice. This can lead artificially high values of R^2 and errors in the coefficients and their interpretation.

Alternative statistical methods are available which solve some of the problems associated with MLRA such as principle component analysis, in which the number of descriptors can exceed the number of observations. This is possible in principle component analysis as the method reduces the dimensions of the model by transforming/condensing the original variables into principle components, which are a set of variables that define the maximum amount of variation in a dataset. Each principle component is orthogonal (and therefore uncorrelated) to the previous principle component of the same dataset. The principle components are constructed so that the first component extracted explains the maximum variance in the dataset. The second component then explains the maximum of the remaining variance of the dataset.

Principle component regression uses these principle components as descriptors in a regression against the dependent variable. However, models created using principle component analysis are often more difficult to interpret that those made with MLRA as each principle component is a combination of a number of factors and the physical meaning can be difficult to define.

Another method available which overcomes the problem associated with MLRA is partial least squares analysis (PLS). In this method the independent variables are transformed into latent variables via linear combinations of the original independent variables, in a similar manner to principle components analysis, although unlike principle components

latent variables are constructed with the intention of maximising their correlation with the dependant variable.

After the latent variables have been constructed they are correlated with the property of interest by MLRA, in which the latent variable are the descriptors. PLS is a particularly powerful tool when the number of independent variables greatly exceeds the number of dependant variables also co-linear descriptors are not a problem.

It can be envisaged that the descriptors, which we propose in 3.1.2, will not correlate highly and the datasets chosen will be sufficiently large that there will always be a minimum of 10 observations to every descriptor and analysis by MLRA will be possible.

2.10 Software

All statistical analyses were performed using JMP version 4.0.2.[8]

2.11 References

1. A.C. Atkinson, *Plots transformation and Regression*. 1985, Oxford: Clarendon Press.
2. S. Chatterjee, B. Price, *Regression Analysis by Example*. 1977, New York: Wiley and Son.
3. R.F. Gunst, R.L. Mason, *Regression Analysis And Its Application*. 1980, New York: Dekker.
4. H.T. Hayslett, *Statistics Made Simple*. 1981, London: Heinemann.
5. A.R. Leach, *Molecular Modelling Principles and Applications*. 1996, London: Prentice Hall.
6. B.F.J. Manly, *Multivariate Statistical Methods*. 1986, London: Chapman Hall.
7. N.A. Weiss, *Elementary Statistics*. 1989, Reading - Massachusetts: Addison Wesley.
8. JMP, *Published by SAS software*. 2000.

3.1 Model Development

3.1 Generating Descriptors

3.1.1 MOLVOL

While many different program and algorithms for the calculation of molecular surface areas are available the program MOLVOL by Dodd and Theodorou[1] was chosen as our method of generating the surface areas from which our descriptors would be calculated. MOLVOL was chosen as it has been seen capable of calculating accurate surface areas and more specifically PSA from previous studies.[2] MOLVOL was also selected, as the source code is publicly available and legal to modify.

MOLVOL calculates the total and individual volume and exposed surface area of a molecule by treating it as an arbitrary collection of fused hard spheres of predefined radii cut by relevant planes.

MOLVOL places the centre of each sphere in positions representing the nuclei. These positions are defined from a set of XYZ coordinates from the input file. The exposed surface area of these spheres is then calculated from radii encoded into MOLVOL. Where two spheres overlap a plane of intersection is drawn perpendicular to the line connecting the two spheres. This plane contains the circle of intersection of the two sphere surfaces. A vector is then assigned that points from the centre of the sphere to the centre of the plane of intersection. It is from this vector that MOLVOL determines which parts of the sphere are exposed (before the plane of intersection) and which are buried (after the plane of intersection). For two spheres intersecting, the area that is buried by either sphere is a dome shaped spherical cap, which is removed by MOLVOL. The parts of the sphere remaining after all of the buried segments have been removed by the planes is described as the "cutout". Spheres that are intersected by two or more other spheres are treated using the same methodology.

The intersecting spheres are then completely decoupled to leave the individual cutout of each individual atom; each of these is given a copy of the intersection planes along with the vector. This form of decoupling means each individual sphere is treated as an

individual problem; this is essential to our study as the proposed descriptors are reliant on the exposed surface area of individual spheres/atoms.

Once the configuration of the all the planes of intersection are known and the molecule has been decoupled into separate atoms, the volume and surface area is determined by treating the cutout as a sum of "cone pyramids" formed by plane intersections and "spherical sectors" formed by the uncut remains of the sphere. The cone pyramids can be thought of as the unexposed part of the cutout and the spherical sector is the exposed part of the cutout.

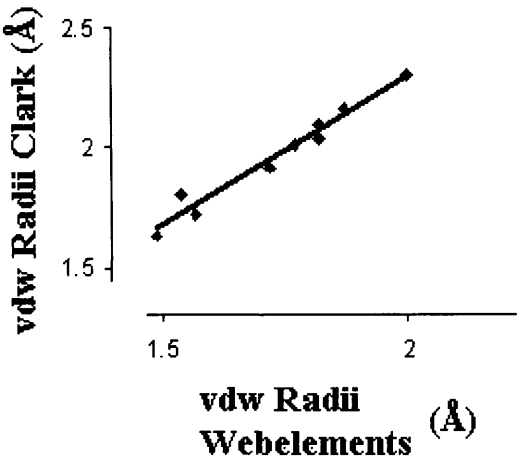
The cone pyramids are defined thus: the base of the cone pyramid is defined as the plane of intersection, in the case of two spheres interacting the base would be a circle and the cone pyramid would be a simple cone. The height of the cone pyramid is the perpendicular distance from the centre of the sphere (nucleus) to the plane of intersection. The planes of intersection which form the base of the cone pyramid are most often intersected with other planes that also cut the sphere, this form of multiple intersection causes the base of the cone sphere to become a complex shape containing numerous plane lines and points of intersection. These complex shapes are referred to as arc-polygons, as they often still possess arcs from the original sphere circle. The surface area of these cone spheres is defined as the convex solid object drawn out by lines connecting the sphere centre with all points on the perimeter of the arc polygon. From the area of the cone pyramid the volume is calculated by taking the product of this area and one-third the height of the cone pyramid. The volumes and surface areas of these regions represent the unexposed surface area and volume of the cutout. Once these volume and surface area are calculated they are removed from the cutout to leave the spherical sector.

To calculate the surface area and volume of the spherical sector, all of the arc polygons are connected to form one or more closed curves on the cutout sphere surface. It is these closed curves that represent the boundary of the exposed surface area. A spherical sector may have more than one series of closed curves on the surface of the sphere, and also more than one contiguous region of uncut sphere. Analysis of the closed curves on the surface of the sphere allows the determination of the exposed surface area of the sphere. The spherical sector surface area is defined as a cone-like solid object whose apex is sphere centre and whose base is the continuous uncut area of the sphere (exposed surface

area). The volume of the spherical sector is calculated as one third the radius of the sphere (the height of the cone) times the surface area of the continuous uncut area.

MOLVOL is reliant on pre-programmed radii to calculate surface areas. These radii were obtained from the study of Clark.[3] and were chosen as they have been seen to be reliable for generating PSA values from previous studies.[2-4] Other radii that were required but not included within those stated by Clark were obtained from web elements.[5] It was seen that radii obtained from the work of Clark and those from web elements were different, although a correlation of 0.966 was seen when the two sets of radii were plotted against each other as shown in fig 3.1.

Figure 3.1: Plot of vdw radii obtained from the study of Clark *vs* radii obtained from web elements



From the fit shown in figure 3.1. The following relation was defined.

$$C = 1.2556W - 0.17 \qquad (3.1)$$

Where C is the value of Clarks radii and W is the value of the radii as obtained from web elements both set of radii are in Å. The radii as obtained from Clark and those extrapolated from web elements are listed in table 3.1.

Table 3.1: van der Waals Radii as implemented into MOLVOL

Atom	Vdw Radii/Å
C	1.90
O	1.74
N	1.82
H	1.50
H attached to O	1.10
H attached to N	1.13
S	2.11
P(II)	2.05
F	1.65
Cl	2.03
Br	2.18
I	2.32
Pt	2.03*
Si	2.46*

*Value as calculated from webelements.

In order for MOLVOL to assign vdw radii an if statement was used, in which radii were assigned from atomic symbol. Atomic numbers were also assigned to each atom to aid determination of molecular fragments. The radii for H attached to O or N were assigned based upon data contained in the connectivity block. The exact nature of this assignment is discussed further in 3.1.4.

In order for MOLVOL to calculate the surface area descriptors required, the first change required was the format and the information contained in the input file of MOLVOL. Figure 3.2 shows the original MOLVOL input file for formaldehyde. The Cartesian coordinates from which MOLVOL calculates surface area contain no information about the chemical environment of individual atoms i.e. bond type, aromaticity or connectivity. For this reason the input file was changed to that of MDL's Molfile format.[6] Molfiles encode the relevant information in the molecule's bond block. Figure 3.3 Contains the MDL molfile for formaldehyde as generated by HYPERCHEM PRO 6[7] and shows how chemical information is encoded in the connectivity block.

Figure 3.2: Original Input file format of formaldehyde program MOLVOL

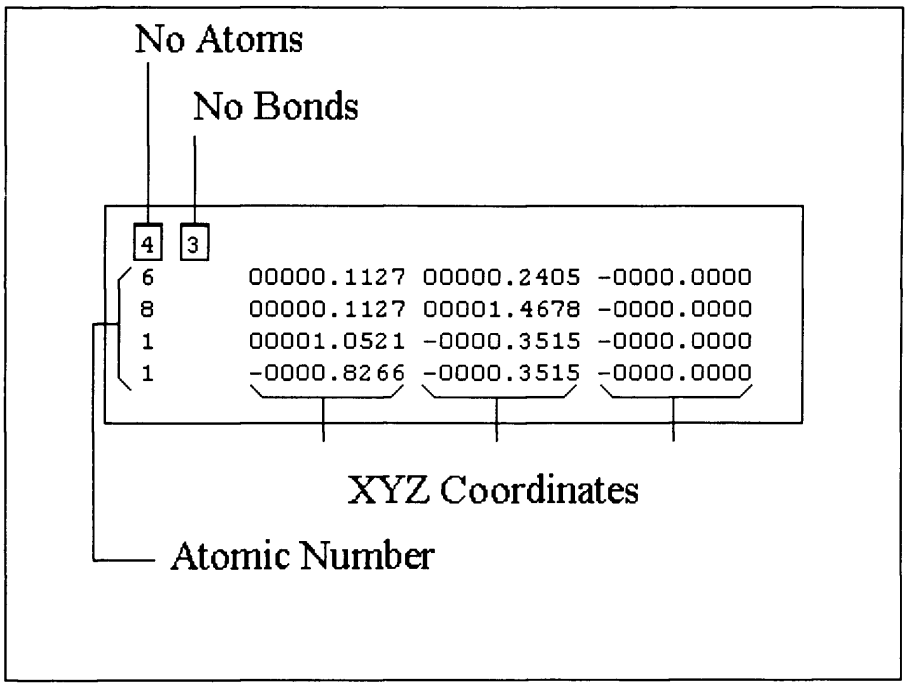
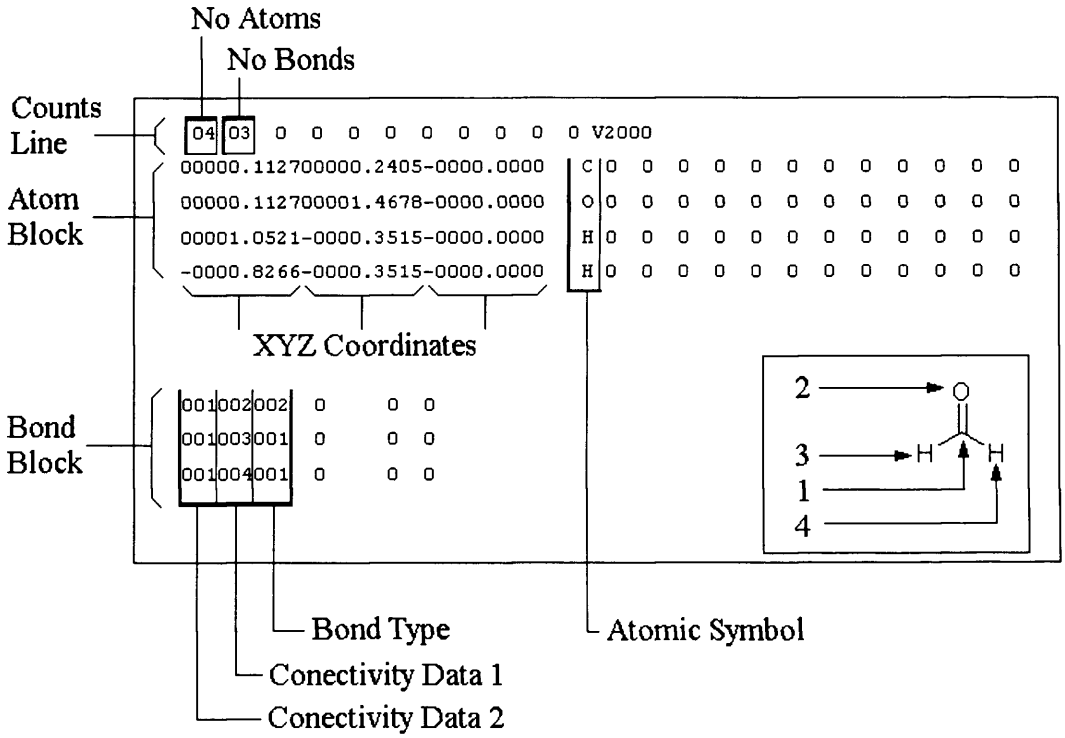


Figure 3.3: MDL Molfile for formaldehyde and connectivity assignment



The connectivity data in the bond block refers to atoms in the atom block. A value of one in the connectivity data refers to line one of the atom block a value of two refers to line two (in the example shown in figure 3.3 the carbon atom and the oxygen respectively).

The first line of the bond block states that atom one is bonded to atom two. The bond type in this example, connecting atom one to atom two has the value two, which indicates that the bond is a double bond. The notation used to define bonds in a molfile is as follows.

1. Single
2. Double
3. Triple
4. Aromatic
5. Single or Double
6. Single or Aromatic
7. Double or Aromatic
8. Any

While further information is encoded within the molfile such as charge and valence in the atom block and stereo information in the bond block, this is not discussed further as it has no relevance to the proposed descriptors.

To aid the speed with which MOLVOL could be used to calculate descriptors, the program was altered so that the name of the desired molfile to be processed could be entered on the command line, and an output file with the extension .out would be produced. As opposed to MOLVOL'S original usage by which a file title Volume.inp would be processed and a file titled volume.out would be produced.

All of the aforementioned changes were made to MOLVOL in the form of a new subroutine entitled Read_dataMol.

3.1.2 Descriptors

As stated in 1.4.5 the use of PSA as a measure of a molecule's hydrogen bonding capacity reduces the various ways a molecule can interact with the environment to a single number. In order to give greater flexibility to PSA as a descriptor a number of surface area descriptors were defined that account for the various ways in which a molecule can interact with its environment.

While it has been shown that PSA represents a molecule's hydrogen bond capacity[8] its traditional definition groups all hydrogen bonding atoms together and does not account for the fact that individual polar atoms in molecules will act as hydrogen bond donors and acceptors. Hydrogen bond donors are defined as hydrogen atoms covalently bonded to electronegative atoms such as oxygen and nitrogen. It is the withdrawing effect of the electronegative atom on the electron density of the hydrogen atom that causes the atom to gain a positive charge while the electronegative atom gains a negative charge. The small size of hydrogen relative to other atoms and molecules results in a large charge density. The partial positive charge on this hydrogen is capabable of interacting with hydrogen bond acceptors, which are hetaroatoms with lone pairs or partial negative charges such as oxygen and nitrogens. Hydrogen bond donors and acceptors are also refered to as hydrogen bond acids and bases.

To account for these properties, PSA was decoupled into its hydrogen bond acid and hydrogen bond base components. The hydrogen bond acid surface area descriptor is denoted as ASA_U and is defined as the total vdw surface area of all hydrogens attached to an oxygen or a nitrogen. The hydrogen bond basicity surface area descriptor (BSA_U) is defined as the total vdw surface area of all oxygen and nitrogen atoms.

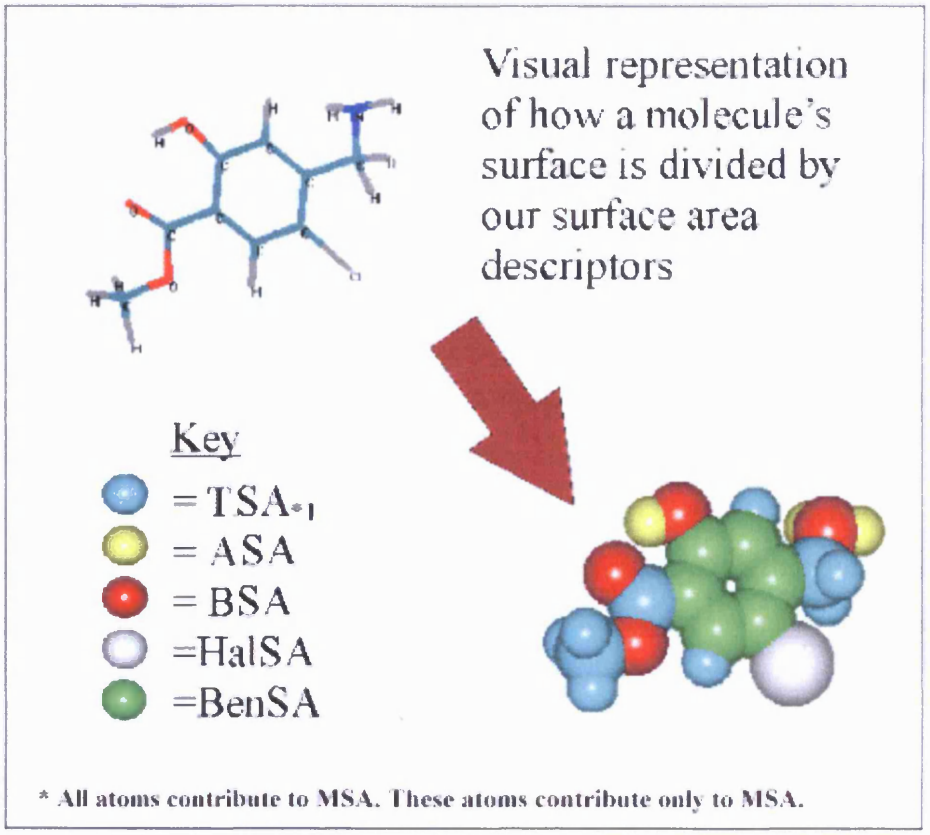
The strength of a hydrogen bond depends greatly on the donor and acceptor. For this reasons two other descriptors ASA_S and BSA_S were defined; these descriptors are intended to scale the hydrogen bond acid and base descriptors to give a more realistic representation of hydrogen bonding. The scaling of these descriptors is discussed further in 3.1.3.

The total surface area of a molecule is also important to its solvation properties. The cavity theory[9-11] of solution states that in order for a solute to be solvated a cavity of suitable size must be created in the solvent. The formation of this cavity is endoergic due to the energy required to disrupt the solvent solvent interactions. The reorganisation of the solvent molecules may also account for a large change in enthalpy and entropy and the introduction of the solvent into this cavity causes various solvent-solute interactions all of which are exoergic. The larger the solute the greater the size of the cavity, and the greater the disruption to the solvent-solvent interactions. Hence total surface area is an apt descriptor for this interaction. The total surface area descriptor (TSA) is defined as the total vdw surface area of the whole molecule.

Simple energy calculations[12] show that the centre of an aromatic ring such as benzene is capable of interacting with hydrogen bond donors and acting as a hydrogen bond acceptor. It has been seen that the hydrogen bond is formed by a small partial charges centered on the ring. The hydrogen bond formed is seen to be about half as strong as a normal hydrogen bond. Via the scaling factors it is possible to include aromatic carbon surface areas into the descriptor BSA_U, although benzene rings tend to interact more strongly with solvents via polar and polarisable effects. For this reason the slight hydrogen bonding basic properties of aromatic carbons is included into BSA_S and also included as a separate descriptor BenSA which is defined as the total surface area of all aromatic carbon. The descriptor is intended to take into account polar and polarisable properties of benzene rings.

The effect of halogen atoms upon solubility is included in the descriptor HalSA, which is defined as the total vdw surface area of all halogen atoms. The HalSA descriptor accounts for the dipole-dipole interaction and induced dipole interactions between the solvent and the halogen atoms of the solute. The partitioning of a molecule by the proposed surface area descriptors is shown in figure 3.4.

Figure 3.4: The partitioning of 4-Aminomethyl-5-chloro-2-hydroxy-benzoic acid methyl ester by the surface area descriptors.



The descriptor TSA was already encoded into MOLVOL and included in the output file. The descriptor ASA_U was assigned based on connectivity data, and calculated as the total exposed surface area of any H attached to an O or N. The BSA_U surface area was assigned based upon atomic number with BSA_U being calculated as the sum of the exposed surface area of any atom with an atomic number of 7 or 8. The definition of the descriptors ASA_S and BSA_S and their calculation is discussed in 3.1.4.

The variety of atoms included in the definitions of ASA_S and BSA_S is wider than that of its unscaled counterparts as the effects of weak hydrogen bonding molecules is accounted for by appropriate scaling. ASA_S is expanded to include the scaled exposed surface area of H attached to alkynes; BSA_S is expanded to incorporate sulphur, phosphorous and carbon in alkanes and alkynes.

The HalSA descriptor was assigned from atomic number. The BenSA descriptor, which is the surface area of any aromatic carbon atoms, was assigned based upon atomic symbol

and the value of bond type in the bond block. MOLVOL was modified so the values calculated for these surface areas were included in the output file.

3.1.3 Scaling Factors

The molecular fragments from which the scaling factors would be assigned were chosen with the intention of keeping the model as simple and general as possible without loss of accuracy. For example hydrogen atoms attached to oxygens were defined as alcohol, phenol and carboxylic acids. Experimentally [13] each type is found to have broadly similar hydrogen bond donor abilities. Similar classifications were made for N-H, oxygen, nitrogen and sulphur. Sulphur has been included in our definition of PSA here as its "slightly polar" properties can be modelled via appropriate scaling. PSA including sulphur have been proven to produce more accurate models than just oxygen and nitrogen based PSA.[14] A series of 49 fragments were defined a list of these fragments is given in table 3.3.

The Abraham A and B values were chosen as the starting point for our scaling. A is the overall hydrogen bond acidity of the solute. The preliminary A^H_2 scale was developed from acid base systems in tetrachloromethane at 298K. From available literature Abraham and co workers [15,16] created the acidity scale denoted K^H_A, this scale fitted all the equilibrium constants onto one single scale irrespective of the reference base. Plots of logK for acids with a given reference base against the logK for acids with another base produced a series of straight lines that intersected at a point. The origin was set to zero for convenience for the resulting A^H_2 scale. As this scale was derived for 1:1 complexation and a more realistic measure of a solutes H bonding acidity as surrounded by solvent molecules, an overall ΣA^H_2 scale was defined, in this scale 0 corresponds to no hydrogen bond acidity and 1 represents a strong monofunctinal acid. For example a value of 1 is expressed for pentachlorophenol if a molecules displays a ΣA^H_2 value of 0.1 its hydrogen bond acidity is one tenth that of pentachlorophenol.

B is the hydrogen bond basicity of the solute. B is derived from Taft's pK^H_B scale [17,18] using a similar methodology as the acidity scale. Taft's pK^H_B is related to the Gibbs free energy of formation of the hydrogen bond complexes in tetrachloromethane at 298K. The scale was generalised to an overall the ΣB^H_2 is a similar manner as that for hydrogen bond

acidity, with a value 0 representing no hydrogen bond basicity and 1 being equal to a strong monofunctinal base such as hexamethylphosphorictriamide (HMPTA). This ability to express free energy related properties in terms of conventional units and the appropriate range of values covered by the ΣA^{H}_{2} and ΣB^{H}_{2} scales makes them an appropriate and apt scale for our scaling factors.

A further strength of the A and B scales is their values are derived from experimental measurements such as changes in the infra red stretching frequency H-X upon formation of a complex B···H-X, gas liquid chromatography data and water solvent and gas/solvent partition data.

Two methods were proposed for the generation of scaling factors for the fragments the first of these was a simple averaging of A and B values while the second method was a more elaborate regression of A and B values against surface area.

3.1.3.1 Averaging of Abraham Scales

Data for experimentally observed A and B Abraham values were collected.[19] From the collected Abraham values, scaling factors for ASA_S and BSA_S were calculated by taking the average of a number of observed A and B values for a specific functional group. An example of this for alcohol is shown in table 3.2.

Table 3.2: Experimentally observed Abraham A and B values for a series of alcohols.

Name	A	B
Methanol	0.43	0.47
Propan-1-ol	0.37	0.48
Butane-1-ol	0.37	0.47
Butane-2-ol	0.33	0.49
Pentane-2-ol	0.33	0.49
2,2-dimethylpropan-1-ol	0.37	0.5
Octanol	0.37	0.48
Decan-1-ol	0.37	0.48
Cyclopropyl carbinol	0.35	0.4
1-Adamantanol	0.32	0.52
Hexafluoroisopropanol	0.77*	0.1*
Pantolactone	0.53	0.55
Average	0.37	0.48

* Value not included in average.

Care was taken in selecting the molecules from which the averages were calculated, no molecules whose A and B values that were outstandingly different to the majority and whose values could be attributed to effects caused by atoms not included in the specific functional group were included in the calculation. In the above example of alcohol, hexafluoroisopropanol was not included in the calculation of the ASA_S and BSA_S scaling values, as its A and B values are much higher than all other values.

Clark[3] stated that evidence from crystal structure surveys and *ab initio* calculations[20-22] indicates that the ether oxygen in an ester is only rarely a hydrogen bond acceptor, and should perhaps be removed from PSA altogether. If the average B value of an ester has a calculated value of 0.45, this is the same B value as reported for a solitary carbonyl. For these reasons the B value of the ether oxygen in the ester is scaled to 0 and the carbonyl oxygen is scaled by the same value as a solitary carbonyl.

The scaling factors were also designed to take in to consideration that certain atoms in functional groups have relativly large hydrogen bonding acid and base properties but small surface areas. For primary, secondary and tertiary amines, Abraham's scale assigns B values of approx 0.48, 0.54 and 0.65 respectively, while surface area as calculated by

MOLVOL is approximately 19, 11 and 5 $Å^2$. The average B value when multiplied by their relevant surface areas produce values that do not correlate with the increase in H-bonding basicity seen in the Abraham scale, due to the relatively small exposed area of the tertiary amine compared to that of the primary amine. In order to resolve this problem the B scaling values for secondary and tertiary amines were increased appropriately.

The initial intention was simply to multiply the calculated atomic surface areas by the appropriate scaling factors as obtained from the Abraham A and B values to give the scaled values of ASA_S and BSA_S. However, this would mean that 'missed' atoms, i.e. those not matched by any defined fragment, would effectively be assigned a value of 1, corresponding to a very strong donor or acceptor. To remedy this, the Abraham scaling factors were tripled. This means that missed atoms would be scaled as weak to medium donors/ acceptors, similar to alcohols (ASA_S) or ether/alcohol (BSA_S).

It has been observed and calculated that intramolecular hydrogen bonding can act to 'tie-up' both acid and base atoms, and reduce hydrogen bond acidity and basicity. While such effects may be represented in the 3D structures via the overlap of vdw surface area of intramolecular hydrogen bonding atoms, it is possible that these effects will not be accounted for accurately enough to fit onto our proposed scales. A series of scaling factors were defined as a contingency if vdw surface area could not account for the reduction in hydrogen bond acidity and basicity. These fragments are defined simply as a H-bond donor sited *ortho-* to an acceptor on an aromatic ring and are shown in table 3.3 Their values were again assigned via averaging of experimentally obtained Abraham values.

It is not as easy to define fragments for hydrogen bonding within aliphatic systems as the number of fragments required would remove a large amount of generality from the model, Also assignment of values to these fragments is difficult as their effects on hydrogen bond acidity and basicity are less pronounced than those around aromatic rings as conformational flexibility in aromatic rings is far more restricted than in aliphatic systems.

3.1.3.2.1 Regression of Abraham Scales

Simply assigning a value to a scaling factor based on its average value in the Abraham scales may not be an accurate way of correcting a surface area to account for H-bonding strength. While accurate values may be obtained for some functional groups this method does not provide evidence, or indicate if key fragments have been missed. Also other atoms like the N in tertiary amines where the scaling factor must be corrected to account for unusually large or small surface areas may have been missed using the previous methodology. For this reason a second method of calculation scaling factors was explored.

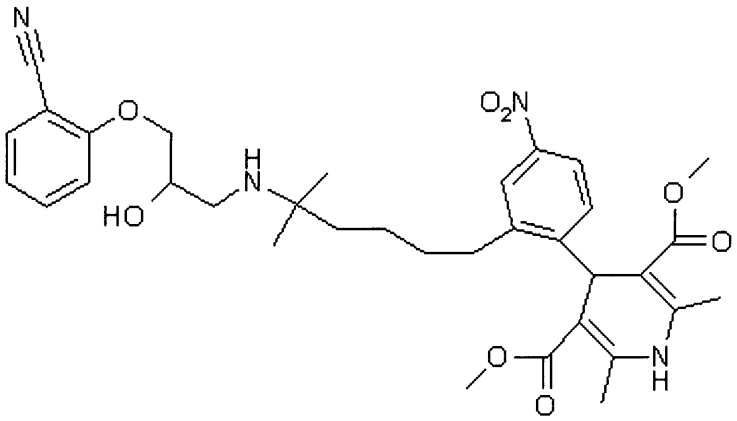
A dataset of 1055 molecules with experimentally observed Abraham values was constructed from the Abraham database[19]. 3D coordinates for each of the molecules in the dataset were obtained from CORINA[23] and energy minimized using AM1 in HYPERCHEM (The justifications for this method of generating 3D structure is given in 3.2.2). A modified version of MOLVOL was used to calculate the total individual surface area for each of our 49 defined fragments for each molecule.

Multiple linear regression analysis (MLRA) of A and B values against the relevant individual surface areas was then performed, i.e. A was regressed against all H attached to oxygen, nitrogen and alkynes. The intercept was removed from the regression so that a molecule with no hydrogen bonding acid or basic surface areas would give A and B values of zero. The coefficients of these regressions were taken as our new scaling factors.

3.1.3.2.2 Results ASA_S

A was regressed against 16 H bond acid surface area fragments. The regression gave an R^2 of 0.82 and an RMSE of 0.126. The residual revealed a number of outliers in the dataset. The first of these outliers is JG-18 (shown in figure 3.5). An observed A value of 2.08 was given this value is very high considering the fragment approach of Platts[24] predicts it to be 0.987. As both theoretical methods substantially over estimate the A value of this molecule we assume the observed value is incorrect and remove it from the dataset. The second largest outlier is 3-bromoacetanilide, comparisons of its reported A value to that of other halogen-substituted acetanilides shows it is seen to be significantly larger than other values, for this reason it was removed from the dataset.

Figure 3.5: Structure of molecule JG18



The molecules mannitol, sucrose and arabitol were also identified as outliers all of these molecules contain large numbers of adjacent alcohols that are capable of interacting via intramolecular H bonding. As fragments are not defined for intra molecular H-bonding in aliphatic systems, and the hydrogen bonding properties of these molecules are governed by such effects we have chosen to remove them from the data set. Chloramphenicol and 2-amino-1-propan-di-tfa were also removed from the dataset as thier A values were seen to be radically different to those of structurally similar molecules.

No individual functional groups were seen to be continually modelled poorly indicating that the 16 fragments assigned to calculate ASA_S were not missing any important functional groups. The regression was repeated with the removal of the outliers. The following statistics were seen R^2 0.88 and RMSE 0.091 a marked improvement over the original regression.

3.1.3.2.3 Results BSA_S

B values were regressed against 32 surface area descriptors for the full set of 1055 molecules and the following statistics were given R^2 0.842 RMSE 0.178.

JG18, chloramphenicol, and N,N diphenlacetamide were reported as outliers. All of these molecules were removed, as their B values were considerably higher or lower than those of structurally similar molecules. N 2,4,6 tetra nitro-N-methylamine was also removed as it contained a fragment (C-N-N-C) that did not occur anywhere else in the data set.

Creating a fragment and scaling factor for this would have been futile, as its value in the regression would be calculated from one point, also the fragment is not common enough in organic molecules to justify its own scaling factor.

The regression was then repeated with the removal of the outliers. The regression yielded the following results an R^2 value of 0.89 and RMSE of 0.143. No individual functional groups were seen to be constantly modelled erroneously indicating that the fragments defined for hydrogen bond bases were comprehensive and that no important functional groups containing these atoms had been missed.

From these regressions of A and B values the coefficients for each fragment were taken as our surface area scaling factors. The coefficient values were all scaled to a range of 0-1 to place them on scale akin to that of the Abraham scales and to aid interpretability. This scaling was performed by scaling the strongest hydrogen bond acid and base to one, the strongest acid was seen to be phenol while tertiary amine was seen to be the strongest base. A list of all these scaling factors is given in table 3.3. This form of scaling makes direct numerical comparisons between the two sets of descriptors difficult as in the averaged obtained scaling factors neither tertiary amines or phenols have values of exactly 1 on the Abraham scale.

Table 3.3: Scaling factors as obtained from regression and averaging method.

N Acids	Regression scaling factor	Averaged scaling factor
1y amine	0.16	0.08
2y amine	0.17	0.08
Aniline	0.22	0.12
Pyrrole	0.69	0.21
Amide	0.41	0.25
Anilide	0.69	0.5
Sulphonamide	0.44	0.45
Thioamide	0.34	0.5
O Acids		
Alcohol	0.46	0.38
Phenol	1	0.54
Carboxilic Acid	0.95	0.6
C Acids		
Alkyne	0.05	0.09
Intra Acids		
Phenol ortho to C=O	0.45	0.05
Phenol ortho to N=O	0.45	0.05
Phenol ortho to O	0.71	0.25
Aniline ortho to O	0.09	0.1
Aniline ortho to C=O/N=O	0.15	0.1

Hydrogen bond bases

N bases		
1y amine	0.21	0.6
2y amine	0.4	1.2
3y amine	1	3
Amide	0.02	0.25
Aniline	0.13	0.4
Cyano	0.06	0.37
Nitro	0	0
Pyridine	0.13	0.75
Pyrrole	0.13	0.25
Sulphonamide	0	0.08

O Bases	Regression scaling factor	Averaged scaling factor
Carbonyl	0.15	0.45
Alcohol	0.16	0.48
Phenol	0.07	0.36
Ether	0.18	0.55
Acid/ester -O-	0	0
Furan/aromatic	0.01	0.15
Nitro	0.01	0.15
Sulphoxide	0.28	0.93
Sulphonamide	0.14	0.36
Sulphone	0.1	0.36
Phosphate	0.18	0.45
Phosphine	0.55	0.45
Intra Bases		
C=O ortho to phenol	0.06	0
N=O ortho to phenol	0.02	0.05
O ortho to Phenol	0.08	0.2
O ortho to aniline	0.04	0.2
N=O ortho to aniline	0.02	0.05
Phospourus		
Phosphate /Phosphine	0.68	0.55
Carbon		
C=C double bond	0.02	0.06
C#C triple bond	0.03	0.13
C Aromatic	0.01	0
Sulphur Bases		
Thiol, sulphide	0.21	0.3

As a final test of the predictive accuracy of the scaling factors obtained from the regression four training and test sets were constructed. 100 molecules were removed from the set the remaining 944 were regressed. The equation produced from this regression was then used to calculate the A and B values for the 100 omitted molecules. This was repeated a further three times with a different randomly selected set of 100 molecules being removed. The results for these models are given in table 3.4. It can be seen that the predictive accuracy of these models is high with an average R^2 of 0.87 for A and 0.85 for B.

Table 3.4: Test set results for A and B regressions.

Test set No.	A R^2	B R^2
1	0.91	0.90
2	0.86	0.81
3	0.85	0.88
4	0.85	0.78
Average	0.87	0.85

3.1.3.3 Scaling factors Discussion

The values shown in Table 3.3 reflect the hydrogen bonding properties of functional groups as determined by many years of careful experiment. As such, there are some useful insights into why the simple definition of PSA is insufficient for our purposes it is also evident that O—H groups are generally stronger acids than are N—H's and Nitrogen bases are usually stronger than their oxygen counterparts.

The scaling factors for hydrogen bond acids are generally higher when obtained by the regression method, although this is due to the scaling of the regression coefficients to place them on a scale of 0-1 where the phenol scaling factor was used to define the top end of the scale, hence direct numerical comparisons between both scales is not possible. There are many broad similarities and trends between the two sets of scaling factors, both sets of scaling factors show that alkynes are relatively weak hydrogen bond acids. Primary and secondary amines are also both seen to be weak H bond acids, the scaling values obtained from averaging state that the same scaling factor can be used for primary and secondary amines, the regression scaling factors are in agreement with this, with a difference of only 0.01 being shown between primary and secondary amines. Anilines are also shown to be relatively weak hydrogen bond acids in both sets of scaling factors. Both show reduction in acidity if an aniline is ortho substituted to either a carbonyl, nitro or ether, The regression scaling factors show that a C=O or N=O will cause less reduction in hydrogen bond acidity than an ether oxygen, while the averaged scaling factors state that C=O, N=O and an ether oxygen will all cause the same reduction in hydrogen bond acidity.

The strongest hydrogen bond acids are seen to be carboxylic acid and phenol, although the regression and averaging method disagree on which of the two is the stronger their values are both relatively very high in both sets of scaling factors.

The largest disagreement between both sets of scaling factors is seen for the intramolecular H bonding fragment for a phenol ortho substituted to C=O or N=O the scaling factors as obtained from averaging suggest that the reduction in hydrogen bond acidity will be heavy and the acidity of the phenol will be reduced to less than that of an alkyne, whereas the regression obtained scaling factors suggest that a much smaller reduction is seen.

For the two scales of hydrogen bond basicity, the difference in numerical values is seen to be even greater than that of the acidity scaling factors. The regression obtained scaling factors were scaled to place them onto a scale of 0-1 by assigning a value of 1 to the strongest base, which was tertiary amine. The large scaling value for tertiary amine means that the scaling factors for regression obtained scaling factors are much lower than those obtained by the averaging method. Although again numerous similarities and trends can be seen between the two set of scaling factors.

Both sets of scaling factors show that the ester oxygen in carboxylic acids and esters, the nitrogen in nitro and aromatic carbon will have negligible or no contribution to hydrogen bond basicity.

The scaling factors obtained from the regression also show the following to be very weak hydrogen bond acceptors: N in sulphonamide, O in furan and O in nitro. The averaged scaling factors disagree with this and state that while these three atoms are weak hydrogen bond acceptors they are not as weak as suggested by the regression scaling factors. Some doubt is cast upon the validity of the regression scaling factors for O in nitro as an increase is seen when comparing the scaling factor for a solitary nitro and that for a nitro ortho substituted to either aniline or a phenol. This increase in hydrogen bond basicity dose not fit in with the experimentally observed decrease in A and B values associated with intramolecular hydrogen bonding. For the averaging obtained scaling factors the expected reduction is seen in scaling values.

The largest scaling factors are seen for tertiary amines. This is not so much a reflection of the strength of tertiary amines as hydrogen bond acceptors but instead a scaling that takes into account the small exposed surface area of tertiary amines. The scaling factors obtained from regression for primary, secondary and tertiary amines show a similar ratio and increase in the series as implemented for the averaged scaling factors to take into account average exposed surface area of these atoms.

The largest difference between the two sets of scaling factors is seen for the N in pyridine, the regression scaling factors state that it should be a weak to medium base while the averaged scaling factors state that it should be a medium to strong base. Experimentally observed Abraham values suggest that the scaling factors obtained from the averaging method are the more accurate and pyridine is a medium to strong hydrogen bond base.

The second largest discrepancy between the two datasets is for the N in a cyano functional group the regression obtained scaling factors show this to have far lower hydrogen bond basicity than that of the averaged scaling factors. This difference is most likely caused by the large average surface area of the N in a cyano (25 $Å^2$) which was not accounted for in the averaging method. The validity and accuracy of these scaling factors with relevance to modelling partition properties is discussed in further detail in 4.2.

3.1.4 Fragments

In order for MOLVOL to assign the hydrogen bonding weighting factors to relevant atoms a subroutine called con_data was added to MOLVOL to classify all the atoms in a molecule via the criteria of our predefined fragments.

The subroutine con_data consisted of a block of do loops and if statements for each of our defined molecular fragments, these blocks of logic statements were named fragment blocks. The fragment blocks were arranged in a specific hierarchy within the program so that no relevant atom would be missed or assigned incorrectly, for example the fragment block for a nitrogen in an amine occurs before the fragment blocks for amides. Which in turn occur before the fragment blocks for intramolecular hydrogen bonding. In this manner a nitrogen is first defined as an amine, as the program continues it checks to see if

the nitrogen is part of an amide, if this is the case the nitrogen is redefined and the original definition is overwritten.

An example of a fragment block for a carboxylic acid is given below in figure 3.6. Figure 3.7 shows the steps through which the subroutine identifies fragments.

Figure 3.6: Fragment block Carboxylic acid.

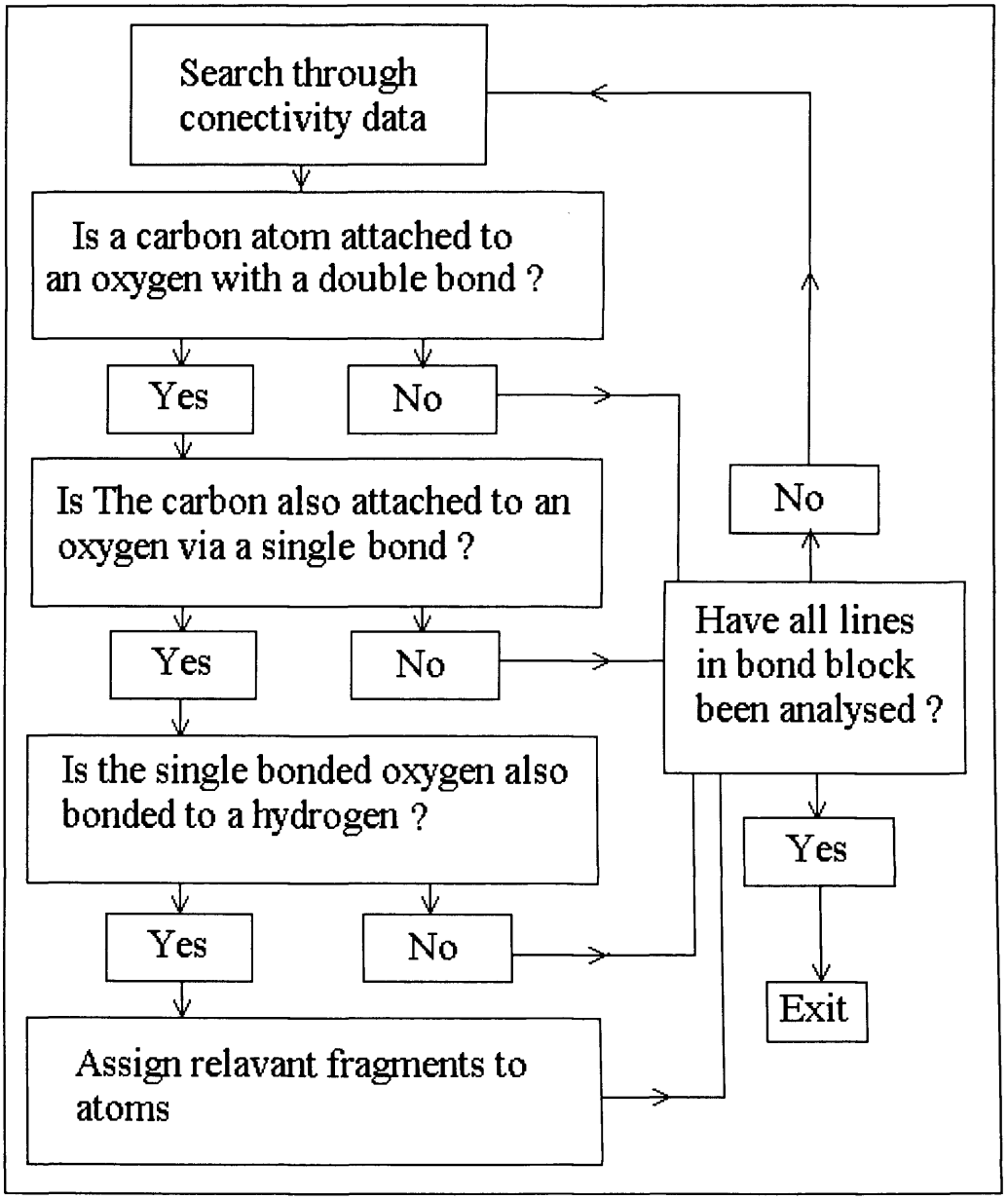
```
C    C=O 3O, C-O-H 4O

   Do J= 0, n_bonds * 2

     If (atnum((con1s(J)-1)).eq.  6 .AND. atnum((con2s(J)-1)). eq.
   >  8 .AND. Btypes(J).eq. 2) then
           Do L= 0, n_bonds * 2
           IF (con1s(L).eq.con1s(J). AND. atname((con2s(L)-1)). eq. '1O' )
   >   then
C    Finding the acidic H within the carboxilic acid
      Do M= 0, n_bonds * 2
           IF (con2s(M).eq.con2s(L). AND. atnum((con1s(M)-1)). eq. 1)
   >  then
      atname((con2s(J)-1)) = '3O'
      atname((con2s(L)-1)) = '4O'
      atname((con1s(M)-1)) = '2H'
         End IF
         End Do
       End If
       End Do
     End If
     End Do
```

The variables and structure of the code as shown above merits discussion in order that anyone wishing to define further fragments would be able to do so easily. The variables in the algorithm are defined as number of bonds is n_bonds and atnum is atomic number. Con1S and con2S are extended versions of the connectivity data as obtained from the molfile. Con1S is an array of all the data from connectivity data 1 followed by all the data from connectivity data 2 obtained from the molfile, con2S is a list of all the data from connectivity data 2 followed by all the data from connectivity data 1. Btypes is an array of all the bond types from the molfile in order twice. An example of this for methanoic acid is shown in table 3.5.

Table 3.5: The arrays con1s, con2s and Btypes for methanoic acid.

Connectivity data1	Connectivity data 2	Bond type	Con1s	Con2s	Btypes
1	2	1	1	2	1
1	4	1	1	4	1
2	3	2	2	3	2
2	5	1	2	5	1
			2	1	1
			4	1	1
			3	2	2
			5	2	1

Figure 3.7: System for finding and assigning carboxylic acid functional group



The need for the extended arrays con1s and con2s is necessary due to the method with which bonds are found within the fragment blocks. For example if one wished to find atom 5 bonded to atom 2, using only arrays of connectivity data, one would first have to search through connectivity data 1 for the occurrence of atom number 5 and then check the corresponding entry of connectivity data 2 for atom number 2. In this case the bond would not be discovered, in order to locate the bond a second search would have to be performed in which connectivity data 1 was searched for atom number 2 and the corresponding value in connectivity data 2 was checked for atom number 5. This need to search connectivity data 1 and 2 individually becomes an increasing problem as the number of atoms within a functional group required to be identified increases.

With the amalgamation of both arrays into one array only one search is necessary as the bond is listed twice once as 2-5 and once as 5-2. The array Btype is also treated in this manner so that the type of bond can be obtained from the array BtypeS.

The if statements that are used to identify specific bonds can be broken down and translated as follows.

IF 1

Fortran 77

```
If (atnum((con1s(J)-1)).eq.  6 .AND. atnum((con2s(J)-1)). eq. 8 .AND.
Btypes(J).eq. 2)
then
```

If the atomic number for the atom in field one of con1s equals six, and the atomic number for the atom in field one of con2s equals eight, and the bond type in BtypeS equals two (a double bond) then this is a carbonyl.

The J in the fortran code is the field in which the specific line of code is analysing if the target bond is not found then the loop adds 1 to the value of J and the next set of fields of con1s and con2s are searched. This is performed until either a match is found or the value of J equals double the number of bonds and all fields of con1s and con2s have been searched.

If a carbonyl is detected then the next if statement would be

IF 2

```
IF (con1s(L).eq.con1s(J). AND. atname((con2s(L)-1)). eq. '1O' )
```

L determines the field that is now searched, J remains the same as in IF 1. So the code would read, if the atom in field one of con1S the same as that determined to be the carbon of a carbonyl? If so then is the atom in con2s an oxygen in an alcohol? If field one is not a match then the value of L is increased by one and field two is checked. Again this is repeated until a match is found or all the fields have been checked. If no match is found after all the fields are searched then the fragment block is exited and the program moves to the next fragment block.

If the program identifies a functional group it is given a label in the array Atname. Atname is an array that contains the information used to identify the relevant atom in the fragment. It is from the values assigned in Atname that the scaling factors are assigned.

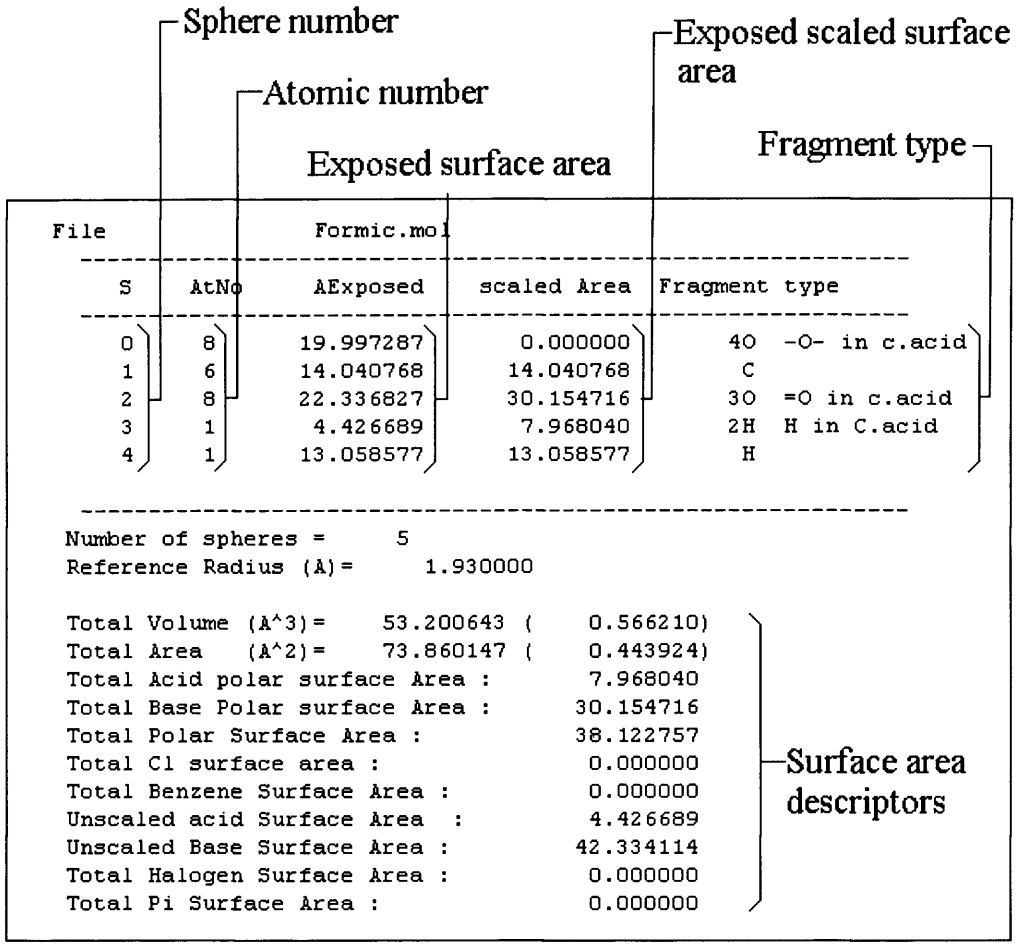
For certain functional groups it is necessary for the fragment block to contain checks to ensure that the same atom was not being counted twice. For example in order to determine if a nitrogen is a primary or secondary amine the number of hydrogen attached must be determined, if the hydrogen in a secondary amine is counted twice then the program would wrongly classify the nitrogen as a primary amine.

3.1.5 Output

The output of MOLVOL was amended to include the calculated surface area descriptors, individual surface area, individual scaled surface area values, atomic numbers and a list of all the fragments in the molecule. The output file was also altered to remove numerous pieces of information that are not required for calculating polar surface area such as CPU time and total volume of sphere.

The output file produced by the modified MOLVOL program of formic acid is shown in figure 3.8.

Figure 3.8: The modified output file from MOLVOL for formic acid



3.1.6 Automation

To aid the speed with which surface area descriptors could be calculated a shell script and an awk script were written to automate the processes of calculating and tabulating surface areas. The shell script named runscale opened a file that contained the names of all the mol files that are intended for calculation. This shell script then calls MOLVOL and processes the first job on the list. After the job is processed the shell script calls an awk script, which collects the surface area descriptors from the output file of MOLVOL and tabulates them in a separate file. With this automation it is possible to calculate surface area descriptors for approximately 50 molecules per second running on a Compaq XP1000 workstation. This rate of generating surface area descriptors is more than adequate for use as a virtual screening tool.

3.2. Generating 3D Structures

3.2.1 CORINA

A necessary requirement for the calculation of the surface area descriptors is a full set of 3-D atomic coordinates. Following the work of others[2] we have chosen to generated these coordinates in a two step process, firstly approximate 3D coordinates are generated using rule and data based computer programs such as CORINA[23], CONCORD[25], MOLGEO[26] or COBRA[27]. These approximate 3D coordinates are then energy minimised to remove any close steric interactions and give more accurate structures.

The program CORINA was selected as the method for generating approximate coordinates as in trials[28] using 639 X-ray structures obtained from the Cambridge Crystallographic Database[21] against six other automated 3D structure generators (CONCORD, ALCOGEN, Chem-X, MOLGEO, and COBRA), CORINA was seen to give a 100% conversion rate and produce structures that most accurately resembled those obtained from the X-ray crystallography. For these trials a structure was determined to be well reproduced if the RMSE deviation of the atomic positions ($RMSE_{XYZ}$) was less than 0.3 Å. Chain geometry was defined as being well reproduced if the RMS deviation of the torsion angles was less than 15°.

CORINA was also seen to remove accurately atom crowding with only 3% of the structures generated containing close contacts, with a molecule being defined as free of close contact interactions if the ratio of the smallest non bonding distance against the smallest acceptable value for this distance was less than 0.8.

While the trial showed that of the six programs CORINA was not the fastest method of generating 3D structures, its rate of conversion of 0.58 s/molecule running on a VAX 6000 computer is rapid enough to meet our needs.

CORINA is capable of producing 3D coordinates from connection tables or a linear code. The linear code method was chosen, as it is a faster method of defining structures and can be performed by hand even for highly complex molecules. The linear code that was used was SMILES[29] (Simplified Molecular Input Line Entry Specification), which is part of the

Daylight tool kit. SMILES are a simple yet comprehensive chemical structure nomenclature. SMILES follow a series of simple rules under which the structure for any molecule can be encoded as a linear string. Examples of these rules are

1. Atoms are represented by atomic symbols.
2. Double bonds and triple bonds are represented by = and #.
3. Branching in the molecule is indicated by the use of branching pairs.
4. Pairs of matching digits indicate ring closure.
5. Lower case letters represents atoms within aromatic systems.

Hydrogen atoms are not defined within SMILES strings and are added implicitly for atoms specified without brackets, from normal valence assumptions. Other rules are encoded into smiles to take into account features such as charge, chirality, and isotopes. Examples of SMILES for a series of organic molecules are shown in table 3.6.

Table 3.6: Example of SMILES

Name	Smiles
Methane	C
Ethane	CC
Ethene	C=C
Acetic acid	CC(=O)O
Benzene	c1ccccc1
Cyclohexanol	C1CCC(O)CC1
Caffeine	Cn1cnc2n(C)c(=O)n(C)c(=O)c12
Nicotine	CN1CCCC1c2cccnc2

3.2.2.1 Optimisation Methods

A method for geometry optimisation or energy minimisation for the approximate 3D coordinates in order to relieve close steric interactions that may have been missed by CORINA had to be identified. Previous studies of PSA as a modelling tool have stated their optimisation method but none have offered any justification behind their choice, Krarup[30] and Clark[2] both used the max2min minimize Tripos force field in SYBYL, Palm[31] chose molecular mechanics calculations using the MM2 force field, while Stenburg[32] used the semi empirical method AM1 .

The selection of an Optimisation method was made using two criteria

1. The structures must be accurate.
2. The method must be rapid enough that structures for large datasets of 100+ molecules can be generated quickly.

Geometry optimisation methods can broadly be classed into three categories molecular mechanics, semi empirical and *ab initio*.

The word "*Ab initio*" is Latin for "from the beginning" meaning that calculations are derived directly from theoretical principles with no inclusion of any experimental data. This usually refers to an approximate quantum mechanical calculation, where the approximations are usually mathematical approximations such as an approximate solution to a differential equation or using a simpler functional form for a function.

Hartree Fock calculations (HF) are the most common form of *ab initio* calculations. In HF calculations two approximations are made. The primary approximation is called the central field approximation. This approximation states that the Coulombic electron – electron repulsion is not specifically accounted for, although its net effect is included in the calculation. The approximate energies calculated in units called Hartrees (1 H = 27.2114 eV) have the exact energy as a lower bound. This approximation means that energies calculated by HF are higher than the exact energy and tend to a limiting value called the Hartree Fock limit.

The second approximation of HF calculations regards the wave function. As the wave function must be described by some functional form and exact functional forms are only known for a few one electron systems, the functions used are most often derived from linear combinations of Slater type orbitals (STO) or Gausian type orbitals (GTO). The wave function is formed from the linear combination of atomic orbitals or basis functions. The exact set of basis functions used is often specified by an abbreviation, such as STO-3G or 3-21G.

The most powerful property of *Ab initio* calculations is that they eventually converge to the exact solution, once all of the approximations have been made sufficiently small.

While these calculations are the most accurate they are also require enormous amounts of computer CPU time, memory and disk space, and so are not really appropriate here.

Semi-empriical calculations are similar to HF calculations except certain factors such as two electron integrals are omitted or approximated. This omission of information is corrected for by the use of curve fitting of appropriate parameters against experimental data to give the best concurrence with the experimental data.

This form of curve fitting causes a problem in semi empirical calculations in that if the molecule being calculated is dissimilar to those in the data base used to parameterise the method, the results produced may be poor, although alternatively if the molecule closely resembles those of the parameterisation set the results may be good.

Semi empirical methods have been more successful in organic chemistry than *ab initio* due to the fact that the limited selection of atoms that occur in organic compounds are well defined and parameterised the molecules studied are rarely large enough to represent a problem.

Molecular mechanics represent the fastest and simplest of all three methods. Their main application lies in calculations on molecules that are too large to be treated with *ab initio* or semi empirical methods such as protein and segments of DNA. For this reason it has become a popular computational method in fields such as biochemistry. The speed with which molecular mechanics are capable of performing calculations is due to the fact that it is totally devoid of any quantum mechanical calculation. Molecular mechanics uses simple algebraic expression to calculate the total energy of the molecule; they are not reliant on factors such as wave functions or electron density. Simple classical equations such as the harmonic oscillator equation are used to describe the energies associated with bond stretching and rotation. All of the constants for these equations are derived from either experimental data or *ab initio* calculations.

In molecular mechanics methods the set of parameters used by the method is referred to as the force field. The choice of compounds when using molecular mechanics methods is fundamental as many force fields are parameterised against very specific classes of

molecule, e.g. proteins. The main failing of molecular mechanics is that many chemical properties are not defined or parameterised such as electronically excited states.

These three methods represent a wide cross section of techniques ranging greatly in accuracy and computational time, with *ab inito* calculations typically taking hours, semi empirical methods taking minutes and molecular mechanics methods taking seconds.

Two separate molecular mechanics force fields were chosen these were AMBER and MM+. AMBER (Assisted Model Building and Energy Refinement) is based on force field developed for computations of protein and nucleic acid molecules. A great deal of development has gone in to the AMBER force field due to its high popularity in academia. AMBER was first designed as a united atom force field [33] and later extended to include an all atom version.[34] As AMBER was developed for the treatment of macromolecules there are few parameters for the treatment of small organic and inorganic molecules. The second molecular mechanics force field selected was MM+,[35] which was developed primarily for small organic molecules.

Two semi empirical methods were also selected; these were PM3 and AM1. AM1 (Austin Model 1) proposed by Dewar *et al*[36,37] is an in improvement of the modified neglect of diatomic overlap (MNDO) method. While AM1 uses the same basic approximations as MNDO, alterations to the functions describing repulsion between atomic cores and assignment of new parameters have significantly improved its performance. The second semi empirical method selected is PM3[38], this method is a reparameterisation of AM1, and differs only from AM1 in the value of the parameters. A much larger number and wider variety of experimental versus computed molecular properties were used to derive the parameters for PM3. Both AM1 and PM3 contain parameters for all atoms commonly found in organic molecules. Neither method includes parameters for transition metals; PM3 is also parameterised for a number of main group elements.

Due to the time consuming and computationally heavy nature of *ab initio* methods only one quantum mechanical Hartree fock method was selected, with basis set 3-21G.

In order to assess the different optimisation methods and their effects upon our descriptors, structures were generated using CORINA and then energy minimized using

the five aforementioned methods for a dataset of 110 organic and drug like molecules. The 3D coordinates acquired directly from CORINA with no form of geometry optimisation were also analysed.

Molecular mechanics and semi empirical calculations were performed in HYPERCHEM running on a 733Mhz PC with optimisation terminating after 2000 cycles or when a gradient of <0.01 kcal mol^{-1} $Å^{-1}$ is attained. *ab initio* calculations were performed using GAUSSIAN98[39] running on a Compaq XP1000 workstation.

To assess the geometrical similarity of the structures produced by each optimisation five molecules were compared (aspirin, dichlofenac, fluoxetine, ibuprofen and papaverine) using the overlay function in HYPERCHEM and RMSE values were generated based on the spatial similarity of the two molecules. These five molecules were selected, as they represent a range of structures similar to those for which our proposed methods will be applied. The structures were also selected as experimental structural data was available, in the form of an X-ray crystal structures obtained form Cambridge crystallographic database.[21] Comparisons between the structures generated from theoretical means offers information about the internal self consistence of these methods, while comparisons of theoretically generated structures to X-ray crystallographic structures offers a method of benchmarking the different theoretical methods. Surface area descriptors were calculated for all six structures for each of the five molecules to give an insight into how structural differences generated by the optimisation method are manifested within the descriptors.

Table 3.7 contains the average RMSE deviations in nuclear positions for five sample molecules. It is clear from these results that optimisation using any of these methods alters the CORINA geometries substantially, leading to large RMSE deviations. As might be expected, the molecular mechanics methods, MM+ and AMBER, agree well with each other, as do the semi-empirical methods, AM1 and PM3. As a test of internal self-consistency, the agreement between methods is encouraging.

Table 3.7: Average RMSE deviations in nuclear positions for five molecules in Å

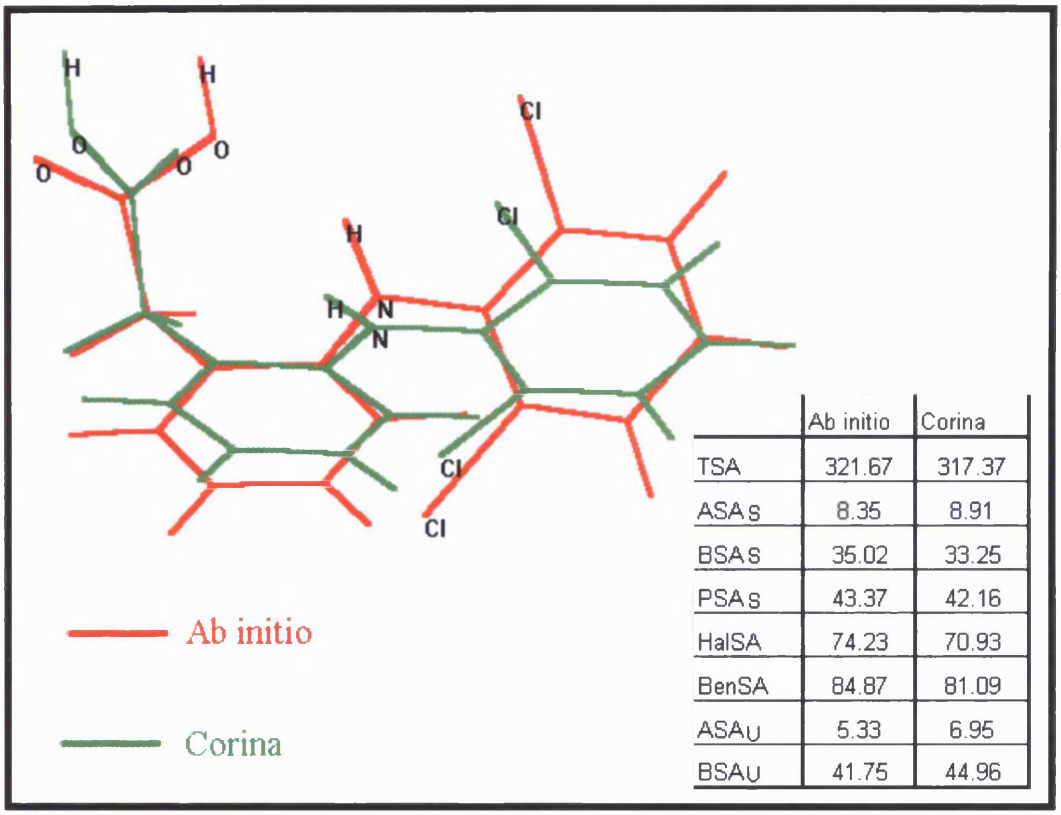
	Crystal Structure	*Ab initio*	PM3	AM1	MM+	Amber
Corina	0.604	0.755	0.612	0.516	0.591	0.803
Amber	0.487	0.775	0.736	0.507	0.296	
MM+	0.387	0.739	0.771	0.547		
AM1	0.354	0.59	0.502			
PM3	0.491	0.861				
Ab initio	0.295					

Comparisons to crystal structures shows as one would expect that the most accurate methods of generating structures is the *ab initio* method, while structures obtained directly from CORINA show the greatest deviation in nuclear position. These results show broadly the expected trends that are associated with optimisation methods with a trade off between time required to run the calculations and the accuracy of the results gained. While the *ab initio* methods are the best the time required for the calculations may be too lengthy to be used as a tool for virtual screening in which extensive libraries of molecules may require processing. The semi empirical method AM1 is seen to have an RMSE value only 0.06 Å higher than that of the *Ab initio* methods, this difference in RMSE is perfectly acceptable when weighted against the speed with which AM1 calculations can be performed. An unexpectedly low RMSE value of 0.387 is reported for MM+ structures against crystal structures.

Although it is expected that simple monofunctional structures can be accurately generated from any of the aforementioned methods, in order to confidently proceed similar overlays were performed in HYPERCHEM for a number of simpler molecules. For straight chain molecules (1-heptanol) and rigid molecules where there is little conformational freedom (2-fluorophenol) the RMSE between CORINA structures and *Ab initio* is seen to be very low (approximately 0.05 Å).

Figure 3.9 shows the overlay of CORINA and *Ab initio* optimised structures for dichlofenac. Distinct differences can be seen in the spatial orientation of the two structures an RMSE of 0.962 Å is seen for the overlaying of the two structures, but these differences are not reflected in the surface area descriptors also shown in Figure 3.9.

Figure 3.9: The overlay of CORINA generated Dichlofenac with its *ab intio* energy minimised analogue and surface area descriptors.

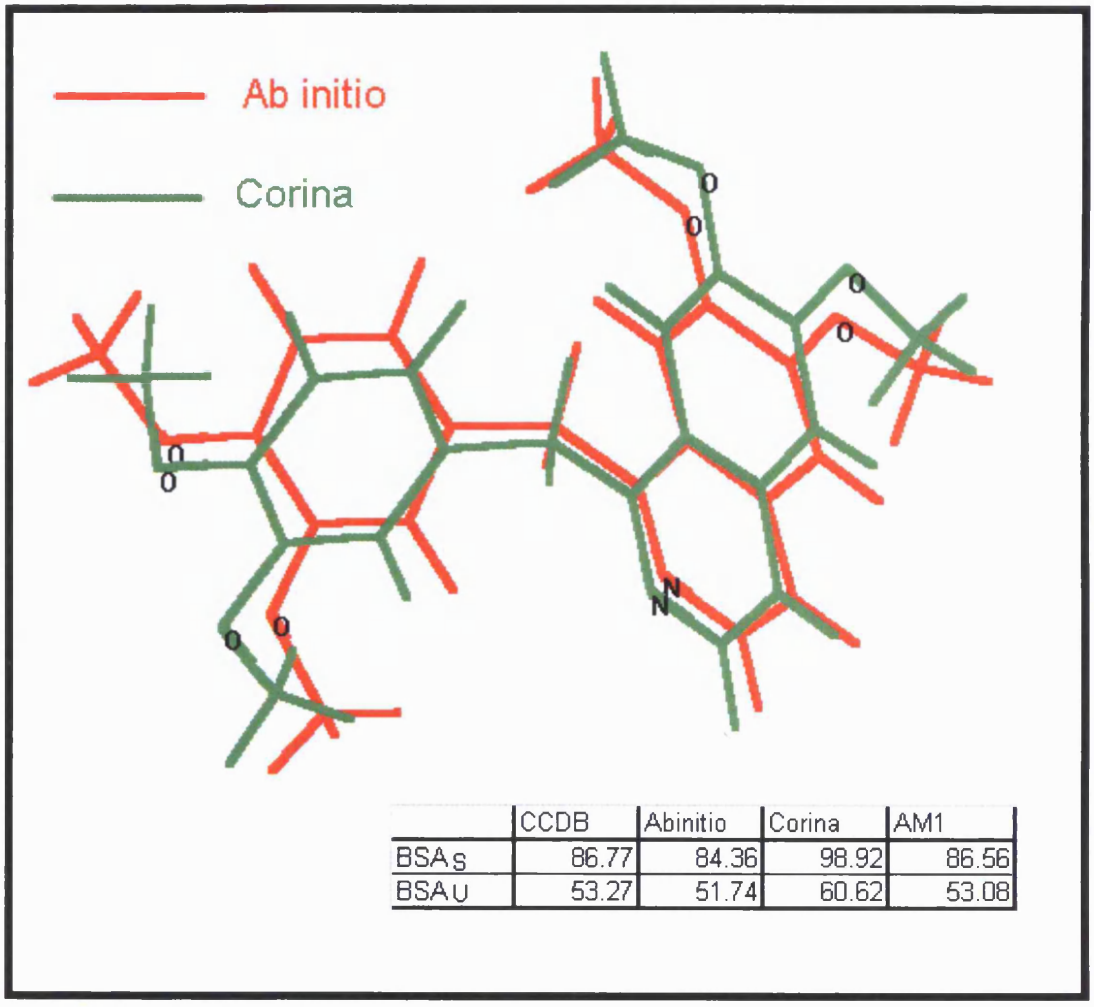


	Ab initio	Corina
TSA	321.67	317.37
ASA S	8.35	8.91
BSA S	35.02	33.25
PSA S	43.37	42.16
HalSA	74.23	70.93
BenSA	84.87	81.09
ASA U	5.33	6.95
BSA U	41.75	44.96

* scaled descriptors are calculated using the method set out in 3.1.3.1

By comparing the overlay diagram of dichlofenac with the descriptors it is clear that while the molecule changes in orientation during energy minimisation the descriptors dont change significantly. This is because for a molecule such as this there is no difference in the two structures that would act to tie up surface area such as the overlapping of vdw radii. However, while it may appear that for molecules were little overlap of vdw radii may occur the choice of optimisation method may be ultimately trivial this is not the case. If we consider a molecule such as papaverine where the same functional group (ether) occurs four times we see the descriptors for BSA_U exhibit a greater difference in values. The overlay diagram and descriptors for Papaverine is given in figure 3.10.

Figure 3.10: The overlay of CORINA generated Paparverine with is *ab intio* energy minimised analogue and surface area descriptors.



	CCDB	Abinitio	Corina	AM1
BSA_S	86.77	84.36	98.92	86.56
BSA_U	53.27	51.74	60.62	53.08

While the two structures shown in Figure 3.10 appear similar in configuration the values for the descriptors BSA_U and BSA_S are overestimated when obtained directly from the CORINA generated structure when compared to descriptors obtained from X-ray crystallography and *ab initio* structures. This overestimation of descriptor values is due accumulative error from the ether oxygens. The *ab initio* and X-ray crystallographic structures suggest this functional group should have a bond angle of 117 – 120° with bond lengths of 1.4 and 1.3 Å. While the structure obtained from CORINA gives similar bond lengths the bond angle for this group are much lower at approximately 106°. This smaller bond angle means that the exposed surface area calculated from the CORINA structure is approximately $3Å^2$ larger than that of X-ray crystal and *ab initio* derived surface areas. It again should be noted that the semi empirical method AM1 is in agreement with the *ab*

initio and X-ray crystallographic descriptors due to its similarity in bond angle for this functional group (116°).

A further method for comparison of the descriptors generated from different methods is possible via multivariate analysis. Surface area descriptors were generated for all 110 molecules as obtained from all structure generation methods. For the TSA descriptor high correlations with R^2 values of > 0.99 are given by all methods. The hydrogen bonding descriptors ASA_U and BSA_U show lower correlation, with ASA giving the lower of the two, both descriptors similar trends to those given in table 3.7 with high correlations being given when similar methods of geometry optimisation are used i.e. descriptors calculated via molecular mechanics methods MM+ and Amber correlate with R^2 value of >0.995 for both ASA_U and BSA_U. The descriptors generated directly from CORINA are seen to correlate the least with all other sets of descriptors with an average correlation of 0.936 for ASA_U and 0.962 BSA_U. The descriptors generated from *ab initio* optimised structures are seen to correlate most highly with descriptors obtained from AM1 optimised structures for both ASA_U and BSA_U with R^2 values of 0.983 and 0.997 being reported respectively. When scaling factors as detailed in 3.1.3.1 are applied to the hydrogen bond descriptors very little difference is seen in the correlations with values remaining the same or changing by small increments of 0.001. The expected trends were again visible with similar methods correlating highly and AM1 correlating more highly with *ab inito* descriptors than any other optimisation method.

This similarity in results obtained from AM1 and *ab initio* methods again indicates that AM1 may represent the best balance of accuracy and time for generating 3D coordinates.

3.2.2.2 Geometry Optimisation Conclusions

From the results it can be seen that as expected the *ab initio* methods are the most capable of producing structures that most resemble closely those obtained from X-ray crystallographic structures. Although due to the time required for *ab initio* calculations and number of optimisations that are required for our models they cannot be seriously considered. The semi empirical method AM1 is seen to be a close second to *ab initio* methods with the average time for calculation being acceptable for our requirements.

The results also showed that for small and inflexible molecules or where few polar atoms are present, the choice of optimisation method can be inconsequential. For more complex molecules such as papaverine the type of optimisation method was more significant, this was reflected in the surface area descriptors as generated from different structures. It was seen that the miscalculation of fundamental structural properties such as bond angles would have detrimental effects on surface area descriptors, a problem that is exacerbated for descriptors such as BSA_S if the angle in question refers to a polar atom. The AM1 generated structures were seen to be very similar to X-ray crystal structures in term of bond lengths and bond angles.

Correlations of the descriptors generated from the different methods showed that the AM1 descriptors most closely resembled those of *ab initio*. Descriptor values obtained directly from CORINA structures correlated the least with those obtained from *ab initio* methods.

From all the evidence it can be concluded that the optimisation step is fundamental to the values generated by surface area descriptors. It can also be concluded that the best method of optimisation is the semi empirical method AM1 as it offers the best trade off between accuracy and the time required to produce structures.

The use of AM1 as the primary method for generating structures is further verified in 4.3 by the construction of individual predictive models based on descriptors obtained from each of the structure generation methods.

3.2.3 Conformational Flexibility and its Effects on PSA

The effect of conformational flexibility on PSA descriptors has been discussed at some length (see 1.4.3) Palm *et al* stated that the PSA_d descriptor which accounts for multiple low energy conformations, would give a better descriptor of molecular surface area than a descriptor that only accounts for one conformation. One difficulty with dynamic measures of PSA is that to conduct a conformational search with energy minimisation for even a small moderately flexible molecule can take several hours of CPU time on a modern workstation.

Studies by Pearlman,[40] Ertl[41] and even Palm[31] have shown that accurate models can be constructed considering only a single conformer. Clark concluded that values for a single conformer generally fall within a few percent of values averaged over many low energy conformations.

The effects of conformational changes upon our proposed PSA descriptors are unknown, and individual descriptors may be found to be more sensitive conformational change, for example ASA_S and ASA_U may be highly effected by intra molecular hydrogen bonding as these would tend to "tie up" the hydrogen bonds donor ability.

Conformational studies

10 flexible molecules were chosen which displayed a good range of functional group types, atom types and conformational flexibility. A detailed conformational search was preformed on each of these molecules using HYPERCHEM. A molecular mechanics method using an MM+ force field was used for the conformation searches as the time taken for semi empirical calculations would have meant that far fewer conformations could be generated.

The conformations search in HYPERCHEM varies the dihedral angle around a specified bond. The method generates random variation of the selected dihedral angles to generate new structures and then energy minimizes each of these. Low-energy unique conformations are stored while high-energy or duplicate structures are discarded. The generation of new starting conformations for the energy minimization uses random variation of dihedral angles. The conformation search was set to perform 1000 optimisations. Each individual conformation was saved as a molfile and passed into our modified version of MOLVOL. Surface area descriptors were calculated for TSA ASA_U BSA_U ASA_S BSA_S HalSA BenSA. The scaling factors for ASA_S and BSA_S those as stated in 3.1.3.1.

An average was taken for each surface area descriptor along with the standard deviation. The average descriptor values along with the descriptor values as calculated from an AM1 optimised structures are given in table 3.8. It should be noted that average values were was not Boltzman weighted as those in the study of Palm.[31,42]

Table 3.8: surface area descriptors as averaged over a number of conformations

Name	Conformations	Rotations		TSA	ASA_S	BSA_S	BenSA	ASA_U	BSA_U	HalSA
Aspirin	39	4	Average	223.01	7.29	54.54	48.16	4.05	66.51	0
			AM1	224.44	7.78	51.94	47.62	4.32	63.56	0
Atropine	89	4	Average	371.35	5.09	94.74	47.78	4.46	48.51	0
			AM1	372.7	5	90.42	48.71	4.38	43.29	0
Dichlofenac	43	3	Average	317.96	8.82	34.67	83.36	8.04	44.54	70.73
			AM1	323.01	8.49	35.48	86.03	6.14	43.48	72.58
Fluoxitine	50	2	Average	377.73	1.03	47.46	92.69	4.29	17.39	59.11
			AM1	383.09	1.03	49.48	96.14	4.28	17.27	55.53
Hexanol	91	4	Average	192.45	4.94	26.92	0	4.33	18.7	0
			AM1	196.04	4.99	27.89	0	4.38	19.37	0
Ibuprofen	18	4	Average	298.7	7.52	27.94	40.73	4.18	39.04	0
			AM1	298.29	7.84	25.65	40.36	4.36	37.91	0
Miconazole	100	4	Average	426.75	0	37.62	95.49	0	23.73	145.31
			AM1	438.78	0	33	102.52	0	21	148.71
O-Nitroaniline	4	2	Average	169.64	2.55	33.82	50.11	7.74	63.69	0
			AM1	169.55	2.21	33.41	50.25	6.54	63.41	0
Papaverine	166	5	Average	423.1	0	80.92	101.01	0	49.6	0
			AM1	417.1	0	86.56	105.92	0	53.08	0
Tetracine	178	5	Average	386.33	1.52	85.52	45.68	4.21	42.31	0
			AM1	391.74	1.5	78.65	47.65	4.16	39.78	0

The values in table 3.8 show that the descriptors calculated from the average of all conformations are very similar in value to those obtained from AM1 optimised structures. The largest differences are given by TSA where an average difference of 4.0 $Å^2$ is seen between the two sets of descriptors, although, this difference is small when we consider that eight of the ten molecules have a TSA value of > 200 $Å^2$ with five of these having surface areas of > 350 $Å^2$. It is interesting to note that for 1-hexanol a molecule, which contains a flexible hydrocarbon tail that the average total surface area for all 91 conformers is very similar to the surface area of the AM1, generated structure.

The hydrogen bond basic descriptors BSA_U and BSA_S show some variance in descriptors, with an average difference of 3.05 and 2.05 $Å^2$ for BSA_S and BSA_U respectively. Tetracine and Papaverine are seen to give the largest differences in BSA_S although this is due in part to the fact that these molecules were given more conformational flexibility than the others and more conformations were generated. Some of the effects of conformational flexibility are removed form the scaled base descriptor as atoms such as the ether oxygen in carboxylic acids and esters are scaled to 0 and hence removed from BSA_S.

The hydrogen bond acid descriptors ASA_U and ASA_S show very little difference in surface area values. This is an interesting result, as it would be expected that conformations would be generated in which intramolecular hydrogen bonds were generated and a resultant reduction in hydrogen bond acid surface area would be observed. For the molecule o-Nitro aniline it would be expected for conformations containing an intramolecular H-bond a decrease in surface area of ASA and BSA to be observed. However over all conformations very little difference was seen in the descriptors with a standard deviation of 0.39 and 0.50 $Å^2$ being reported for ASA_S and BSA_S.

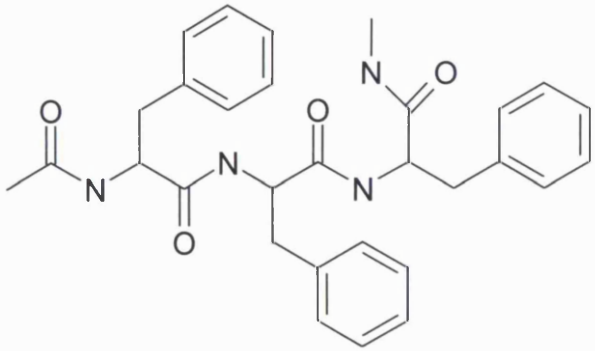
While these results suggest that conformational searches are unnecessary and that the surface area descriptors generated from a single conformation of low energy are almost equal they indicate that the descriptors may not encode certain 3D information such as intramolecular hydrogen bonds. The use of these conformationally averaged descriptors as a predictive tool is assessed further in 4.5.

3.2.4 Encoding 3D Information

As was stated earlier, we expect intramolecular hydrogen bonding to tie up polar surface area and cause descriptors such as ASA and BSA to be reduced. Conformational studies suggest that for molecules where an intra molecular H-bond may occur there is no reduction in ASA and BSA. It can also be hypothesised that if the descriptors are not representing intramolecular hydrogen bonding via overlap of vdw radii then other 3D effects we would wish to model such as cavity effects are also not being accounted for within the descriptors.

In order to assess this problem, conformational searches were performed on the peptide triphenylaniline (Phe-Phe-Phe) again using an MM+ force field in HYPERCHEM. The structure of triphenylaniline is shown in figure 3.11. This peptide was selected as it has potential to display intra molecular H-bonds and is capable of forming a wide range of conformers with diverse range 3D shapes. The search produced 134 different conformations. Surface area descriptors were calculated using MOLVOL.

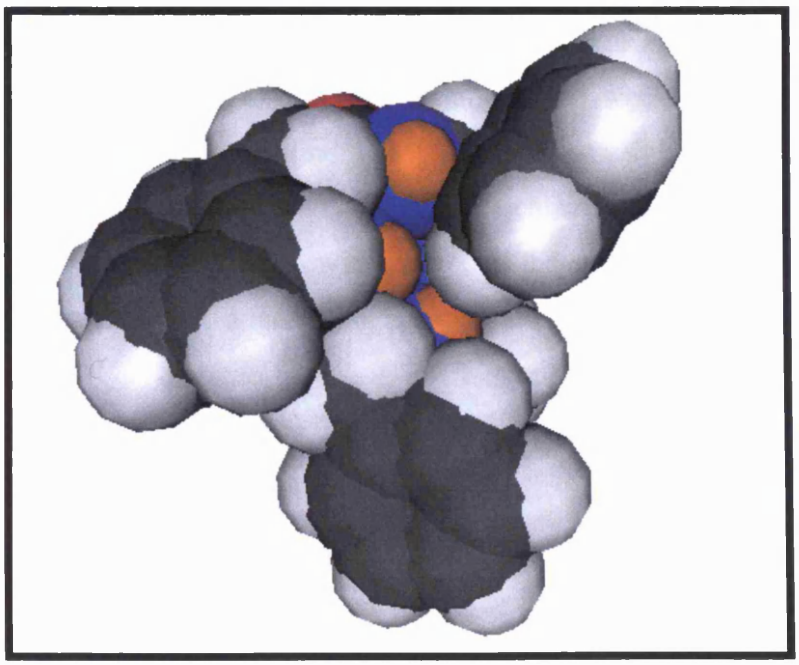
Figure 3.11: Structure of Triphenylaniline



The surface areas generated had a very small range of values 17.94 $Å^2$ – 21.17 $Å^2$ for ASA_U and 101.40 $Å^2$ – 120.11 $Å^2$ for BSA_U. This range of values is low when we consider the number of polar groups in the molecule and the fact that the geometrical variation in the conformations was high. Five of the conformations contained intramolecular H-bonding causing helical structures to occur with polar groups being locked on the inside of the molecule, but for these molecules ASA and BSA were not lower than straight conformations without intramolecular H-bonding.

Figure 3.12 shows a van der Waals surface for a conformation of peptide I where an intramolecular H bond occurs. The vdw Radii used in Figure 3.12 are the same as those stated by Clark[3] and the same as those used in MOLVOL to calculate the descriptors.

Figure 3.12: van der Waals surface of Triphenylaniline



The orange spheres in figure 3.12 represent hydrogen attached to Nitrogen (blue spheres) of amides. The surface areas of these hydrogens are approximately $3Å^2$, a standard value for a hydrogen of this type. Two of these hydrogens will be unable to interact with any solvent molecule due to the steric effects caused by the benzene rings. This phenomenon may cause our descriptors to over predict hydrogen bond acidity. A solution to this is to consider only the solvent accessible surface area.

3.2.4.1 Solvent Accessible Surface Area

As stated in 1.4.2 many different methods have been proposed for the calculation of solvent accessible surface area (SASA). We have chosen to assess two separate definitions of solvent accessible surfaces obtained from two different methods. From these surface descriptors were defined and scaled using the methods stated in 3.1.2 and 3.1.3.1.

The first solvent accessible surface area method is that of Lee and Richards. This method was originally designed to quantify the effects of protein folding and the burial of hydrophobic side chains. The accessible surface of a molecule is defined as the van der Waals envelope of the molecule expanded by the radius of the solvent sphere about each atom centre.[43]

The second method of generating SASA is that of Fraczkiewicz *et al*[44] as implemented in the program FANTOM[45] and available as the web service GETAREA.[46] In this method SASA is calculated by finding solvent-exposed vertices of intersecting atoms, this method avoids calculating buried vertices, which are not needed to determine the accessible surface area. GETAREA was selected as the contribution of each individual atom towards SASA is reported a factor that is essential to the calculation of descriptors.

3.2.4.2 Expanded van der Waals Radii

The vdw radii defined in MOLVOL for each atom was increased by 0.7 Å, thus causing any gap between two van der Waals surfaces that is less than 1.4 Å (the radius of a water molecule) to be filled.

For triphenylaniline the variation in descriptors calculated over all conformations was much higher than that of the previous approach, with values of ASA_U ranging from 7.1 – 26.68 $Å^2$, giving the descriptors a spread of values four times greater than that of the previous study. BSA_U surface areas gave a range of values from 92.22 $Å^2$ – 145.5 $Å^2$ a spread of values three times that of the previous study. For the five conformations where intramolecular hydrogen bonding had been identified significantly lower values for ASA_U were reported. Where an intramolecular H-bond was present an average surface area of 13.77 $Å^2$ was found, where there was no intra molecular H-Bond an average ASA_U of 19.96 $Å^2$ was found. This same pattern was not seen for the hydrogen bond basic descriptors with both types of conformation giving a BSA_U value of approximately 120 $Å^2$. A list of the surface area descriptors as obtained from a number of conformations is given in table 3.9. The conformations adopted for molecules with intramolecular H-bonding in this study arrange themselves so that two of the nitrogens in the peptide are locked on the inside of the molecule and the oxygen (red spheres) of the carbonyls faces outward exposing themselves fully to the solvent. Figure 3.13 shows a van der Waals surface for the same conformation of triphenylaniline give in figure 3.12 using the expanded radii. The cavity that contained the hydrogen attached to the nitrogen is now filled causing two of the hydrogen surface areas to be reduced to 0.0 $Å^2$, causing a significant reduction in ASA_U to 7.1 $Å^2$. The nitrogen locked inside the molecules no longer contributes to BSA with their exposed surface areas being reduced to zero. Similarly as expected these results are reflected in the scaled descriptors ASA_S and BSA_S (shown in table 3.9).

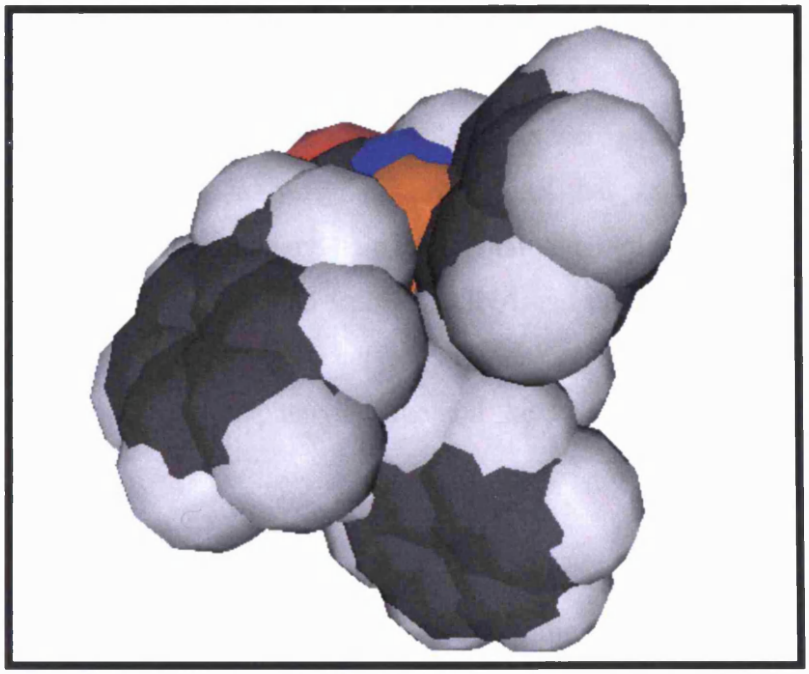
Table 3.9: Average hydrogen bond descriptors for Peptide I for conformations with and without intramolecular Hydrogen bonding

Radii	Intra molecular H bond	TSA* Å^2	ASAs* Å^2	BSAs* Å^2	BenSA*Å^2
Original	YES	604.76	15.16	129.36	145.13
	NO	595.35	15.31	125.83	139.31
Plus 0.7 Å	YES	669.66	10.12	146	131.48
	NO	657.21	14.97	141.41	117.22

*Surface areas are averaged over all conformers

Where the radii suggested by Clark are applied the ASA_S descriptor show no discrimination between conformers where polar atoms are buried due to helical structures caused by intramolcular H-bonding and those without, inflation of the radii by 0.7 Å gives a difference 4 Å^2. Although similar reductions are not seen in BSA_S for either set of radii this is due to the reduction in BSA from the burial of N atoms in intramolecular hydrogen bonding conformations being counteracted by the increased exposure of the oxygens which are forced to the outside of the molecule.

Figure 3.13: Expanded van der Waals surface area of Triphenylaniline



The expanded radii were also applied to the four conformations generated for o-Nitro aniline generated in 3.2.3. One of the four conformers was seen to contain an

intramolecular H-bond between the aniline and the nitro group. Using our original radii all four conformations were seen to give very similar descriptor values with the intramolecular H bonding conformer reporting only a 7% decrease in ASA_U compared to the average value on the non-intramolecular hydrogen bonding conformers. When the radii are expanded by 0.7 Å a larger decrease is seen in ASA_U for the intramolecular hydrogen bonding conformer with its value being 14 % smaller than the average value of the non intramolecuar hydrogen bonding conformations. This reduction is in keeping with experimentally observed A values for O-nitro aniline where a reduction of 13% is seen when compared to aniline.

3.2.4.3 GETAREA

A FORTRAN program was written to process the output files of GETAREA so that descriptors analogous to those stated in 3.1.2 could be generated. The descriptor HalSA could not be calculated due to the restricted number of atoms that are encoded into GETAREA. The program was also designed to scale the SASA calculated by GETAREA to account for hydrogen bonding strength using the scaling factors stated in 3.1.3.1 were applied. The vdw radii stated in table 3.1 were used in GETAREA and a probe radius of 1.4 Å was selected.

Solvent accessible surface area descriptors were generated for 96 simple organic and drug like molecules. This was the same dataset that was used in the optimisation studies of 3.2.2 but excluding any halogen containing molecules. The 3D coordinates for these molecules were obtained from CORINA and optimised using the semi empirical method AM1. Solvent accessible surface area descriptors were also calculated for a number of conformations of triphenylaniline and the four conformations of o-nitroaniline generated in 3.2.3.

For triphenylaniline the descriptors showed a decrease in the value of ASA_S when the geometry of the conformer causes acidic hydrogen to be buried inside a cavity (conformation shown in figure 3.12). The individual buried acidic hydrogen gave a SASA of 0 Å and hence offered no contribution to ASA_U or ASA_S. As with the expanded radii

method, the descriptor BSA_S and BSA_U were not seen to be as strongly influenced by the conformational flexibility of the peptide.

The four conformations of o-Nitroaniline showed a reduction in ASA_U for the conformation in which an intramolecular hydrogen bond was observed. The descriptors for the four conformations are given in table 3.10.

Table 3.10: SASA descriptors calculated for four conformations of o-nitroaniline.

Intra molecular Hydrogen bond	No	No	No	Yes
TSA	308.91	308.91	306	305.5
ASA_U	47.92	47.91	46.44	17.11
BSA_U	103.95	103.95	103.37	96.1
BenSA	109.95	109.88	110.47	107.52

Slight reduction is seen in BSA_U for the intramolecular hydrogen bonding conformation although this is much less significant than the reduction in ASA_U. The size term TSA is seen to be unchanged over the conformations.

For the dataset of 102 simple organic and drug like molecules comparisons were made between descriptors obtained using the original radii stated in 3.1.1, expanded radii and SASA. This comparison was performed using multivariate analysis the results of the analysis for the descriptors TSA, ASA_U and BSA_U are given in table 3.11.

Table 3.11: Multivariate analysis (R^2 values) of descriptors obtained using three different surface area methods.

TSA	Expanded Radii	Original Radii
SASA	0.965	0.964
Original Radii	0.999	

ASAu	Expanded Radii	Original Radii
SASA	0.888	0.899
Original Radii	0.934	

BSAu	Expanded Radii	Original Radii
SASA	0.943	0.921
Original Radii	0.967	

The correlations show familiar trends, of all descriptors TSA gives the highest correlations values indicating that all three methods give similar results for TSA. The

highest correlation for TSA is seen between the descriptors obtained using the original radii and those obtained using the expanded radii, these two methods are seen to give the best correlation for ASA_U and BSA_U. It is not unexpected that these two methods show the best correlation, as they are both calculated using essentially the same algorithm. The high correlations for TSA suggest that both descriptors are encoding virtually the same information about molecular size. SASA descriptors are also seen to correlate very highly with exposed surface area descriptors, combined with the findings for the conformations of o-nitro aniline where TSA was seen to be unaffected by conformational change and molecular geometry we can conclude that for simple organic and drug like molecules there is a strong relation between total solvent accessible and total exposed surface area, and as a descriptor it is relatively insensitive to conformational change.

The hydrogen bonding descriptors ASA_U and BSA_U show more variation that TSA with ASA_U giving the lowest correlations in agreement with previous results. It is surprising that the SASA descriptors show marginally higher correlation with exposed surface area descriptors calculated using the original radii, as previous findings have show that descriptors calculated with the original radii do not account for 3D information.

The similarities between solvent accessible and exposed surface descriptors is due to the simple organic and drug like molecules used in the dataset not being complex enough in 3D structure to cause cavities and bury atoms which would be ignored by the exposed surface area descriptors.

3.2.5 Conformations and SASA Conclusions

It has been seen that if descriptors are generated for exposed surface area using the radii stated in 3.1.1, the effects of conformational change on these descriptors is so slight that descriptors calculated from a single conformation of lowest energy fall within a few percent of those obtained by taking an average of all conformations obtained from a detailed search.

Via inspection of the descriptors it was seen that this immunity to the effects of conformational change is caused by the descriptors failing to identify specific 3D

properties such as the overlap of vdw radii upon formation of intramolecular hydrogen bonds and the burial of atoms inside the molecule where they cannot interact with solvent. The application of two separate methods for calculation of solvent accessible surface produced descriptors, which reflected more realistically the effects caused by molecular geometry. The descriptors calculated from these solvent accessible surface area descriptors showed that TSA was largely unaffected by conformational change while ASA and BSA were seen to be more sensitive.

Through the calculation of descriptors using three different methods for a large dataset of simple organic and drug like molecules it was seen that solvent accessible descriptors and exposed surface area descriptors were similar for simple organic and drug like molecules. This similarity can be attributed to the simple organic molecules not containing any cavity effects and shielding of atoms by steric effects. Effects such as intra molecular hydrogen bonding which are common in simple organic molecules account for some of the variance between the two types of descriptor although these could be accounted for by predefined fragments which will act to reduce the exposed surface area appropriately during the scaling process.

If the exposed surface area descriptors calculated using our original defined radii are capable of accurately modelling partition processes we can hypothesis that these descriptors can be accurately obtained from a single conformation of lowest energy. This hypothesis is further tested in chapter 4 where various different models of partition processes are constructed using descriptors obtained from a variety of surface areas.

3.3 References

1. L.R. Dodd, D.N. Theodorou, *Mol. Phys.*, **1991**. 72. 1313.
2. D.E. Clark, *J. Pharm. Sci.*, **1999**. 88. 815.
3. D.E. Clark, *J. Pharm. Sci.*, **1999**. 88. 807.
4. K. Palm, P. Stenberg, K. Luthman, P. Artursson, *Pharm. Res.*, **1997**. 14. 568.
5. Webelements, *website address - www.webelements.com*.
6. Molfiles, *Website address - http://www.mdl.com/downloads/public/ctfile/ctfile.pdf*. 2003.
7. HyperChem, 6., *Published by Hypercube.inc.* 2000.
8. P. Stenberg, U. Norinder, K. Luthman, P. Artursson, *J. Med. Chem*, **2001**. 44. 1927.
9. M.H. Abraham, J. Liszi, *J. Chem. Soc., Faraday trans,I*, **1978**. 74. 1604.
10. R.A. Pierotti, *Chem. Rev.*, **1976**. 76. 717.
11. H.H. Uhlig, *J. Phys. Chem.*, **1937**. 41. 1215.
12. M.F. Levitt, M.F. Pertutz, *J. Mol. Biol.*, **1988**. 201. 751.
13. M.H. Abraham, *Chem. Soc. Rev*, **1993**. 22. 73.
14. S. Winiwarter, N.M. Bonham, F. Ax, A. Hallberg, H. Lennernas, A. Karlen, *J. Med. Chem*, **1998**. 41. 4939.
15. M.H. Abraham, P.P. Duce, P.L. Grellier, D.V. Prior, J.J. Morris, P.J. Taylor, *Tetrahedron Lett.*, **1988**. 29. 1587.
16. M.H. Abraham, P.L. Grellier, D.V. Prior, P.P. Duce, J.J. Morris, P.J. Taylor, *J. Chem. Soc. Perkin Trans.2*, **1989**. 1587.
17. R.W. Taft, *Steric effects in Organic Chemistry*, ed. M.S. Newman. 1956, New York: Wiley.
18. J. Shorter, *Pure. Appl. Chem.*, **1994**. 66. 2451.
19. M.H. Abraham, *Descriptor Database*, **1066**.
20. H.J. Bohm, S. Brode, U. Hesse, G. Klebe, *Chem.-Eur. J.*, **1996**. 2. 1509.
21. I.J. Bruno, J.C. Cole, J.P.M. Lommerse, R.S. Rowland, R. Taylor, M.L. Verdonk, *J. Comput.-Aided Mol. Des.*, **1997**. 11. 525.
22. P.R. Rablen, J.W. Lockman, W.L. Jorgensen, *J. Phys. Chem. A*, **1998**. 102. 3782.
23. J. Gasteiger, J. Rudolph, J. Sadowski, *Tetrahedron Comp*, **1990**. 3. 31.
24. J.A. Platts, D. Butina, M.H. Abraham, A. Hersey, *J. Chem. Inf. Comput. Sci.*, **1999**. 39. 835.
25. R.S. Pearlman, *Chem. Des. Auto.*, **1987**. 1. 1.
26. E.V. Gordeeva, A.R. Katritzky, V.V. Shcherbukhin, N.S. Zefirov, *J. Chem. Inf. Comput. Sci.*, **1993**. 33. 102.
27. A.R. Leach, K. Prout, *J. Comput. Chem.*, **1987**. 11. 1193.
28. J. Sadowski, J. Gasteiger, G.J. Klebe, *J. Chem. Inf. Comput. Sci.*, **1994**. 34. 1000.
29. D.J. Weininger, *J. Chem. Inf. Comput. Sci.*, **1988**. 28. 31.
30. L.H. Krarup, I.T. Christensen, L. Hovgaard, S. Frokjaer, *Pharm. Res.*, **1998**. 15. 972.
31. K. Palm, K. Luthman, A.L. Ungell, G. Strandlund, P. Artursson, *J. Pharm. Sci.*, **1996**. 85. 32.
32. P. Stenburg, K. Luthman, H. Ellens, C.P. Lee, P.L. Smith, A. Lago, J.D. Elliott, P. Artursson, *Pharm. Res.*, **1999**. 16. 1520.
33. S.J. Weiner, P.A. Kollman, D.A. case, U.C. Singh, C. Ghio, G. Alagona, S. Profeta, P. Weiner, *J. Amer. Chem. Soc*, **1984**. 106. 765.

34. S.J. Weiner, P.A. Kollman, D.T. Nguyen, D.A. Case, *J. Comp. Chemistry*, **1986**. 7. 230.
35. N.L. Allinger, *J. Am. Chem. Soc*, **1977**. 99. 8127.
36. M.J.S. Dewar, E.G. Zoebisch, E.F. Healy, J.J.P. Stewart, *J. Am. Chem. Soc*, **1985**. 107. 3902.
37. M.J.S. Dewar, K.M. Dieter, *J. Am. Chem. Soc*, **1986**. 108. 8075.
38. J.J.P. Stewart, *J. Comp. Aided Mol. Design*, **1990**. 4. 1.
39. M.J. Frisch, *Gaussian*, **1998**.
40. R.S. Pearlman, *Physical Chemical Properties of Drugs.*, **1980**. 10. 321.
41. P. Ertl, B. Rohde, P. Selzer, *J. Med. Chem.*, **2000**. 43. 3714.
42. K. Palm, K. Luthman, A.L. Ungell, G. Strandlund, F. Beigi, P. Lundahl, P. Artursson, *J. Med. Chem.*, **1998**. 41. 5382.
43. B. Lee, F.M. Richards, *J. Mol. Biol.*, **1971**. 55. 379.
44. R. Fraczkiewicz, W. Braun, *J. Comp. Chem.*, **1998**. 19. 319.
45. Fantom, *N. Oezguen, R. Fraciekicz, University of Texas Medical Branch*. 1998.
46. GetArea, *Web Site address* - http://www.scsb.utmb.edu/cgi-bin/get_a_form.tcl.

Chapter 4. Partition models

4.1 Data

A dataset of 110 molecules with experimentally determined values of water/octanol, water/chloroform, and water/cyclo-hexane partition coefficients (denoted $logP_{oct}$, $logP_{CHCl}$, and $logP_{cyc}$ throughout) was compiled. The data was collated from two sources, namely Zissimos *et al*'s LSER study of partition coefficients[1] and the MedChem02 database.[2] The molecules fall into two broad classes, being either simple organic molecules or more complex 'drug-like' molecules. These molecules represent a good cross section of the molecules for which our methods will be applied in later studies.

The solvent systems used were chosen as they cover a range of interaction types: both octanol and water are H-bond acids and bases, albeit of different strengths, while chloroform is an acid but not a base, and cyclohexane is neither. Further, they form three-quarters of the 'critical quartet' of partitions proposed by Leahy *et al*[3], designed to encode all important interactions for solvation (insufficient data was available for the final solvent of the quartet, propylene glycol dipelargonate, to be included). This is therefore a stringent test of descriptors and models. Molecules were chosen to represent a range of both chemical and numerical diversity, with maximum and minimum values for $logP_{oct}$ of 5.40 and –1.09, $logP_{CHCl}$ 6.21 and –2.00, and $logP_{cyc}$ 5.24 and –4.88. This dataset of 110 molecules is the same as that used in 3.2.2, 3.2.3 and 3.2.4.

4.2.1 Descripors and Scaling Factors

For this initial study 3D molecular structures were generated using CORINA,[4] and were subsequently optimised using AM1, as implemented in HyperChem 6,[5] with an optimisation criterion of <0.01 kCal mol^{-1} $Å^{-1}$ following the conclusions of 3.2.2.1. futher analysis of optimisation methods applied to models is given in 4.3.

All surface area properties were calculated from these AM1 geometries using our locally modified version of MOLVOL. Two sets of scaled descriptors were calculated using the appropriate scaling factors. These two sets were the set obtained by averaging observed

Abraham values defined in 3.1.3.1 and those obtained from regression of observed Abraham values defined 3.1.3.2.

To test the quality of the modified PSA models, comparisons were made against models derived from Abraham's LSER descriptors, as calculated by the group contribution approach.[6] All models, whether based on PSA or LSER, were found with multivariate linear regression analysis (MLRA) using JMP Discovery software.[7]

4.2.2 Results and Discussion

Table 4.1 contains the results of the initial attempts to model partition coefficients using PSA-type descriptors. Single parameter fits, using TSA, PSA_U *etc.*, are very poor indeed, with typical R^2 values of 0.05 – 0.15, and hence are not considered further. The simplest model in Table 4.1, employing just total and unscaled polar surface areas, is clearly unsatisfactory for all three solvents, typically accounting for only 50-60% of the variance in the data and giving RMSE errors almost twice those from LSER models. Thus, it seems that simple PSA-type descriptors are incapable of forming accurate models of these partition processes. The 'completeness' of the dataset, at least in terms of physical properties spanned, is confirmed by the fact that the LSER model of $logP_{oct}$ is not significantly different to that recently published for 8200 compounds in the logP* list of the MedChem97 database.[8]

Table 4.1: Partition models using unscaled surface area descriptors

Model	$logP_{oct}$			$logP_{CHCl}$			$logP_{cyc}$		
	R^2	RMSE	F	R^2	RMSE	F	R^2	RMSE	F
TSA + PSA_U	0.577	0.792	75.7	0.614	0.941	88.1	0.542	1.16	65.7
TSA + ASA_U + BSA_U	0.578	0.794	50.2	0.714	0.836	88.2	0.618	1.065	59.2
TSA + PSA_U + HalSA + BenSA	0.75	0.613	82	0.642	0.94	47.1	0.583	1.117	38.1
TSA + ASA_U + BSA_U + HalSA + BenSA	0.751	0.616	65.1	0.758	0.776	65.1	0.675	0.991	44.9
LSER*	0.906	0.378	208.1	0.874	0.544	150.4	0.854	0.663	126.9

* calculated in the manner of ref, 6

Breaking down PSA into acid and base surface areas yields slightly improved statistics in two cases, $logP_{CHCl}$ and $logP_{cyc}$, but no significant change for $logP_{oct}$. Closer inspection reveals that PSA_U is dominated by H-bond base atoms (PSA_U and BSA_U are correlated with $r = 0.99$) such that they are effectively interchangeable, this high corelation of PSA_U

and BSA_U means that no model should contain both descriptors. It is well known[7] that $logP_{oct}$ has no dependence on H-bond acidity, so either descriptor can be used equally well here and ASA_U is not significant in the model. On the other hand, both $logP_{CHCl}$ and $logP_{cyc}$ are strongly affected by H-bond acidity, since water is a stronger H-bond base than either solvent, such that the extra flexibility afforded by this model is significant. Table 4.1 demonstrates that including the surface areas of halogen atoms and aromatic carbons (HalSA and BenSA) improves the quality of fit substantially, due mainly to the ability of the two descriptors to encode important polarisability properties of a molecule which are neglected by PSA-type descriptors. In each case, the most flexible model explains around 15% more of the data than models produced using only TSA and PSA_U, which highlights the importance of polarity/polarisability as well as size and H-bonding. Once again, decomposing PSA gives no improvement for $logP_{oct}$, but results in markedly better statistics for the other two solvents. These results demonstrate that it is possible to improve upon 'standard' PSA_U simply by summing the surface areas of different atom types, rather than just those expected to be involved in hydrogen bonding.

Despite these improvements, it is still evident that the models do not take into account all the factors that determine partition coefficients, since even the best results are 12-15% less accurate than the equivalent LSER models.

The first set of scaling factors that were applied were those obtained from averageing of Abraham Values. Applying these scaling factors to the calculation of PSA_S, ASA_S, and BSA_S results in remarkable improvements in modelling all three partitions, as reported in Table 4.2. Considering $logP_{oct}$ first, Table 4.2 shows that even the simplest model, employing just molecular size and scaled PSA, is a great improvement over the unscaled equivalent, explaining 18% more of the variance in $logP_{oct}$ and reducing the RMSE error by 0.18 log units. The form of this model is also encouraging, showing that molecular size increases $logP_{oct}$ while polarity/H-bonding reduces it. Increasing the flexibility of the model by breaking down scaled PSA into its component parts yields an improvement of around 0.05 in R^2, unlike in the unscaled models above. Adding in the halogen and aromatic surface areas improves statistics still further, such that the final 5-parameter model has $R^2 = 0.85$ and RMSE = 0.48, improvements of 0.27 and 0.31 over the original model. To highlight this improvement, Figure 4.1 shows observed vs. calculated values of $logP_{oct}$ for both, along with the analogous comparisons for $logP_{CHCl}$ and $logP_{cyc}$.

Table 4.2: Partiton models made using scaled surface area descriptors

Model	$logP_{oct}$			$logP_{CHCL}$			$logP_{cyc}$		
	R^2	RMSE	F	R^2	RMSE	F	R^2	RMSE	F
TSA + PSA	0.754	0.61	164	0.776	0.734	186.11	0.786	0.809	196.81
TSA + ASA_S + BSA_S	0.807	0.542	147.89	0.891	0.516	288.29	0.855	0.67	207.71
TSA + PSA + HalSA + BenSA	0.826	0.518	124.18	0.786	0.727	96.136	0.787	0.816	96.96
TSA + ASA_S + BSA_S + HalSA + BenSA	0.851	0.481	118.72	0.898	0.5	184.78	0.858	0.669	125.43
LSER*	0.906	0.378	208.1	0.874	0.544	150.4	0.854	0.663	126.9

* calculated in the manner of ref, 6

Figure 4.1: Observed vs. calculated values $logP_{oct}$, $logP_{CHCL}$, and $logP_{cyc}$, of oct, chl and cyc using TSA+PSA (a,c,e) and full scaling (b,d,f)

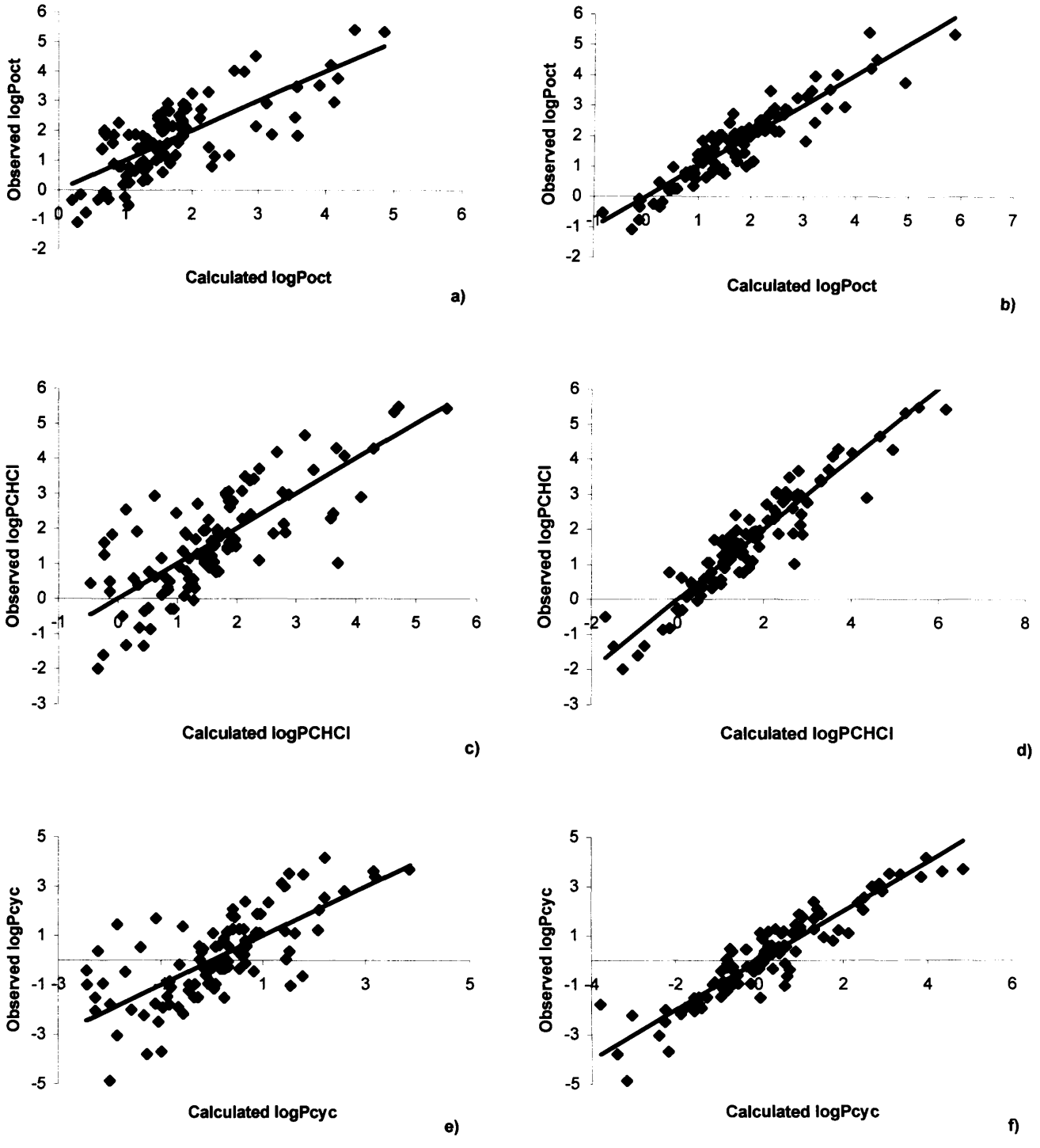


Table 4.2 also shows that the best surface area models are not as accurate as the LSER, and therefore not as accurate as many of the dedicated algorithms for calculation of $logP_{oct}$, such as ClogP. However, it should be emphasised that the purpose of this work is not to generate yet another $logP_{oct}$ calculator, but to use this and other water-solvent

partition data as a test of PSA and related descriptors. In this context, the accuracy attained and improvements made here are, sufficient to justify the modifications made.

A similar pattern of improvements is obtained for $\log P_{CHCl}$, though here the improvement in statistics due to scaling PSA is even greater: the simplest scaled model is slightly better than unscaled one in Table 4.1, with R^2 increased by 0.15 and RMSE reduced by 0.21 log units. In this case splitting up PSA gives substantially greater accuracy, but including more descriptors is barely significant. Nonetheless, the 5-parameter model is the most accurate found in this study, and is actually better than the LSER model, with almost 90% of variance explained and an RMSE of just 6% of the total spread of data.

The final partition coefficient, $\log P_{cyc}$, is the most difficult to model of the three, since the two solvents are possibly the most dissimilar imaginable, and hence covers the largest range of values (over 10 log units) considered here. This is reflected in Table 4.1, in which the statistics for $\log P_{cyc}$ are the poorest. It is encouraging, therefore, that the results in Table 4.2 represent a substantial increase in accuracy; here, the scaled 2-parameter model is 0.24 better in R^2 and 0.35 log units better in RMSE. As with $\log P_{chl}$, splitting up PSA improves results still further but inclusion of HalSA and BenSA is less useful, and again the 5-parameter surface area model is slightly more accurate than its LSER equivalent. In contrast with $\log P_{oct}$, very few methods for the rapid prediction of these, or indeed many other partition coefficients directly from structure have been reported, and the results in Table 4.2 indicate that these models are among the most accurate yet developed.

The scaling factors obtained from the regresion method stated in 3.1.3.2. were applied to the models, the results are shown in Table 4.3.

Table 4.3: Partition models made using scaled surface area descriptors obtained from regresion.

Model	$logP_{oct}$			$logP_{CHCl}$			$logP_{cyc}$		
	R^2	RMSE	F	R^2	RMSE	F	R^2	RMSE	F
TSA + PSA_S	0.621	0.761	86.567	0.849	0.592	298.59	0.777	0.8226	184.79
TSA + ASA_S + BSA_S	0.856	0.4699	208.94	0.85	0.592	198.6	0.794	0.794	135.13
TSA + PSA_S + HalSA + BenSA	0.74	0.634	74.239	0.859	0.576	0.576	0.789	0.807	97.314
TSA + ASA_S + BSA_S + HalSA + BenSA	0.868	0.454	135.73	0.86	0.578	126.87	0.796	0.798	80.414
LSER[a]	0.906	0.378	208.1	0.874	0.544	150.4	0.854	0.663	126.9

It is clear that there is very little difference between the two sets of scaling factors when the results of Table 4.2 and 4.3 are compared. If the five parameter models from each of the two models are compared it is seen that only $logP_{oct}$ shows improvement when the regression scaling factors are applied and this is only an improvement of 0.01% in R^2 and a decrease of 0.02 in RMSE. $LogP_{CHCl}$ and $logP_{cyc}$ both show a decrease in R^2 and an increase in RMSE. A possible explanation for this pattern is that $logP_{oct}$ is not influenced heavily by hydrogen bond acidity. This suggests that it is the hydrogen bond acidity term in the regression scaling factors that causes the models to be inferior to those calculated using the averaged scaling factors.

The source of error within the regression from which the ASA_S scaling factors were calculated may be due to the fact that of the 1055 molecules in the regression all contained hydrogen bond basic atoms under our definitions but 562 contain no hydrogen bond acidic molecules and had an A value of zero giving the regression significantly less data to train the scaling factors.

While the models produced are very similar, the improvement given by the model of $logP_{oct}$ when the regresion scaling factors are applied is outweighed by the loss of accuracy in the models of $logP_{CHCl}$ and $logP_{cyc}$. For this reason we determine the averaged scaling factors to be those that will be applied in all future models.

4.2.3 Significance of descriptors

The t-ratio and coefficients for five parameter models scaled using the averaged scaling factors are given in Table 4.4. These values reveal that size (TSA) is generally the most important term, followed closely by base and acid surface areas. The largest cross-correlation between descriptors is just r = 0.47 for this dataset, such that MLRA is appropriate and direct interpretation of coefficients is meaningful. The models broadly follow expected trends, with the size term always increasing logP and H-bonding decreasing it. Further, ASA_S is less significant for $logP_{oct}$ than the other partitions due to the similar basicity of water and *n*-octanol. Halogen and benzene surface areas play a lesser role, typically acting to increase logP, although perhaps surprisingly HalSA is not statistically significant in a model of $logP_{CHCl}$. Thus, not only are the models developed statistically valid, but they also reflect known physicochemical properties of the solvent systems. Using the critiria that any t-ratio of less than 2 is insignificant within the model the HalSA decriptor can be removed from all the proposed models and BenSA can be removed from $logP_{cyc}$, Although it can be stated that HalSA may not be insignificant to these partitions for a larger dataset that contained a higher number of halogen molecules.

Table 4.4: Coefficients and T-ratios of models of $logP_{oct}$, $logP_{CHCl}$ and $logP_{cyc}$

	TSA	ASA_s	BSA_s	HalSA	BenSA
$logP_{oct}$	0.015 (13.6)	0.033 (2.19)	-0.035 (-9.6)	0.005 (-1.8)	0.01 (5.0)
$logP_{CHCl}$	0.02 (17.1)	-0.225 (-14.4)	-0.044 (-11.6)	-0.004 (-1.3)	0.006 (2.6)
$logP_{cyc}$	0.021 (14)	-0.227 (-11)	-0.66 (-13.1)	0.002 (0.52)	0.004 (1.4)

[a] t-ratios given in parenthesis.

4.2.3 Predictive Accuracy

R^2_{CV} values of 0.806, 0.88, and 0.82, respectively, indicate that the models reported here are capable of making reliable predictions. However, a more realistic test of predictive ability lies in the construction of training and test sets. Five randomly selected test sets of 22 data points (20% of the full dataset) were removed from the dataset, models were then built on the remaining data and used to predict logPs for the omitted molecules. an average variation of 10% was found, but when averaged over all test sets the accuracy is

been shown to have advantages over the method used by Platts for platinum complexes.[24,25] Bowdery showed that the structures calculated by the new method were indeed more accurate and in better agreement with experimental data, and so optimisations at the mPW1PW/LanL2DZ level for all other molecules within our datasets. These structures were obtained as XYZ coordinates and converted to molfiles using HYPERCHEM PRO 6. It was from these molfiles that surface area descriptors were generated using MOLVOL. MOLVOL was modified to calculate an additional surface area descriptor, PtSA that would contain the exposed surface area of the Pt atom. Descriptors were scaled using the methods detailed in 4.9.

4.10.4.1 $logP_{oct}$ Results

Using scaled and un-scaled surface area descriptors a number of models were constructed for all datasets, the results of which are given below in table 4.8.

Table 4.8: $LogP_{oct}$ Surface area models of 24 platinum containing complexes

Number	Descriptors	R^2	RMSE	F-ratio
1	PSA_u TSA	0.85	0.412	59.207
2	PSA_u TSA PtSA	0.93	0.29	90.87
3	PSA_S TSA PtSA	0.85	0.419	38.204
4	PSA_S TSA PtSA HalSA BenSA	0.85	0.448	19.871
5	TSA PtSA HalSA BenSA ASA_u BSA_u	0.94	0.291	43.413
6	TSA PtSA HalSA BenSA ASA_S BSA_S	0.85	0.49	15.723

The results of table 4.8 show the unscaled polar surface area descriptors give better models of PSA than those where scaling factors are applied. This is not an unexpected result as the scaling factors were designed to model the effects of hydrogen bonding in organic not inorganic molecules. We can hypothesise that an atom's distance from the metal core will affect the accuracy of our scaling factors. Atoms directly bonded to the metal core will show the most difference in H-bonding to that of organic molecules while functional groups located at a sufficient distance from the metal will be expected to behave more similarly to those found in organic molecules.

The R^2 and RMSE values show that model number 5 has the highest R^2 value although it is only marginally better than the three-parameter model (number 2) and RMSE is seen to be equal. For this reason we determine that model 2 is the best as the increase in R^2 by 1% by the inclusion of three other descriptors is not acceptable. The t-ratios for model number five are given in table 4.9.

Table 4.9: t-ratio values for $logP_{oct}$ model of 24 platinum containing complexes*

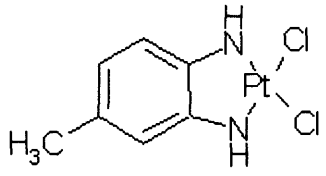
Descriptor	t-Ratio
Intercept	-4.45
TSA	9.25
ASA_U	-0.23
BSA_U	-5.24
HalSA	0.30
BenSA	-0.90
PtSA	2.76

*Model number 5 in table 4.8.

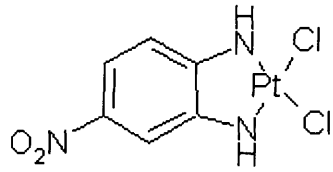
The t-ratios reveal that the descriptors HalSA, BenSA and ASA_U have little significance, hence the similarity between models 2 and 5. Multivariate analysis reveals that BSA_U and PSA_U correlate with an R^2 value of 0.995 indicating that both descriptors are encoding virtually the same chemical information; this high correlation combined with the insignificance of the ASA_U descriptor explains why the splitting of PSA yields no improvement for this dataset.

For a model of $logP_{oct}$ such as this it would be expected that a molecule's hydrogen bond acidity would not be as important as its basicity, which would act to hinder partition into octanol. The large positive TSA t-ratio value is also expected for $logP_{oct}$, acting to encourage larger molecule in to the octanol phase. While the conclusions gleaned from the t-ratios are in agreement with the physical properties of the octanol/water partition, their reliability is questionable due to the restricted size of the dataset. Therefore the dataset was expanded to include 15 more platinum containing complexes.

FIGURE 4.6: Structures of Dianil 8 and Dianil 2



Dianil2



Dianil8

Initial analysis of the expanded dataset showed the molecule dianil 8 (shown in figure 4.6) to be an outlier. Its reported $logP_{oct}$ value of 0.5 is very high when compared to structurally similar molecules within the dataset. The potential error of this value can be highlighted further by comparison of $logP_{oct}$ values of toluene and nitrobenzene, where substitution of a methyl group for a nitro group on an aromatic ring causes the expected decrease on $logP_{oct}$, in contrast dianil2 and dianil8 show an increase in $logP_{oct}$ when a nitro group is substituted for a methyl group. For these reasons dianil 8 was removed from the regression. Table 4.10 shows the results from the regressions of a wide variety of descriptors.

Table 4.10: $LogP_{oct}$ Surface area models of 39 platinum containing complexes

Number	Descriptors	R^2	RMSE	f-ratio
7	PSA_U TSA	0.833	0.608	89.60
8	PSA_U TSA PtSA	0.860	0.502	93.73
9	PSA_S TSA PtSA	0.714	0.806	29.24
10	TSA ASA_U BSA_U PtSA	0.896	0.494	73.13
11	TSA ASA_S BSA_S PtSA	0.725	0.800	22.47
12	PSA TSA PtSA HalSA BenSA	0.87	0.559	44.38
13	TSA PtSA HalSA BenSA ASA_U BSA_U	0.918	0.453	59.31
14	TSA PtSA HalSA BenSA ASA_S BSA_S	0.877	0.553	38.09

The models show similar trends to those of the dataset of 24 $logP_{oct}$ values, with the scaling of the PSA descriptor lowering the quality of the models. Splitting of PSA into its components only marginally improves the fit.

By comparison of the four-parameter models numbers 10 and 11 it is seen that scaling of the descriptors reduces R^2 by 17%. A similarly large reduction is seen when the three-parameter models 8 and 9 are compared, both of these reductions are much larger than those given by models constructed from the small datasets. This is in concordance with our previous hypothesis that the scaling factors are not suitable for representing the H-bonding abilities of inorganic molecules. The expanded dataset contains a wider range of functional groups, which offers a wider source of potential errors for the scaling factors.

When the scaling factors are applied to a six parameter models, the reduction in accuracy is far smaller than for models with less descriptors. When the t-ratios of models 13 and 14 are compared we see that in the scaled model the HalSA and BenSA descriptor have a much higher significance and that the model is far more dependent upon these descriptors, this increased dependency can be attributed again to the poorer quality of the scaled surface areas.

Similarly to the small dataset the best model for this dataset is the six parameter un-scaled model although again this model is only a small improvement over the three parameter model and that the improvement in accuracy is not merited by the inclusion of three extra descriptors. The t-ratios for the six parameter model are given in table 4.11.

Table 4.11: T-ratio values for $logP_{oct}$ model of 39 platinum containing complexes*

Descriptor	T-ratio
Intercept	-6.793
TSA	12.720
ASA_U	1.312
BSA_U	-3.743
HalSA	2.149
BenSA	0.503
PtSA	4.641

* Model number 13 in table 4.10

The t-ratios again reflect the properties expected for octanol water partition with TSA and BSA_U being the most dominant and having opposing effects upon $logP_{oct}$. The ASA_U descriptor is now seen to give a small positive value, which is in agreement with our previous surface area model of $logP_{oct}$ for organic molecules (see 4.2.2). The low t-ratio

values of ASA_U and BenSA indicate that they can be removed from a model of this dataset.

4.10.4.2 $logP_{CHCl}$ Results

Following the same methodologies used for the $logP_{oct}$ datasets a selection of models was constructed using a variety of descriptors. The results of these regressions are given in table 4.12.

Table 4.12: $LogP_{CHCl}$ Surface area models of 14 platinum containing complexes

Number	Descriptors	R^2	RMSE	f-ratio
14	PSA_U TSA	0.93	0.64	70.81
15	PSA_U TSA PtSA	0.93	0.67	42.94
16	PSA_S TSA PtSA	0.93	0.69	41.18
17	TSA ASA_U BSA_U PtSA	0.93	0.71	29.42
18	TSA ASA_S BSA_S PtSA	0.93	0.72	28.43
19	PSA TSA PtSA HalSA BenSA	0.95	0.65	28.07
20	TSA PtSA HalSA BenSA ASAu BSAu	0.96	0.61	27.20
21	TSA PtSA HalSA BenSA ASA_S BSA_S	0.96	0.57	31.14

From the data it can be seen that even a simple model of PSA_U and TSA is capable of modelling $logP_{CHCl}$ and that a simple model of PSA_U TSA is the best model for this dataset as it uses the least descriptors. Unlike the previous models of $logP_{oct}$, the application of the scaling factors has no effect on the models. This difference in the models between scaled and unscaled descriptors is due to the dataset containing a very narrow range of functional groups. Analysis of the scaled and un-scaled ASA and BSA descriptors illustrates how similar these descriptors are, with ASA and ASA_U correlating with 0.90 and BSA and BSA_U correlating to within 0.96. It should be noted that the descriptors in the former models of $logP_{oct}$ do not correlate highly enough for this to be of concern.

A more worrying concern caused by the small number of functional groups within this dataset is the heavy correlation it causes between the hydrogen bond acidic and basic descriptors, with ASA and BSA correlating with 0.932 and BSA_U and PSA_U correlating with 0.923. Therefore any model in which PSA is decoupled in table 4.12 should be

discarded due to the errors associated with internal correlation of descriptors. For these reasons the best model in table 4.12 is number 19, the t-ratios for model 19 are given below in table 4.13.

Table 4.13: T-ratio values for $logP_{oct}$ model of 15 platinum containing complexes*

	t-ratio
Intercept	-1.622
TSA	8.723
HalSA	-1.826
BenSA	1.644
PtSA	-1.704
PSA_U	-8.559

While these t-ratios are sensible and in fitting with our previous studies of $logP_{oct}$ their validity should be taken with caution as the dataset is very limited.

4.10.5 Combining Organic and Inorganic Datasets

While the above models all show the potential of surface area methods in prediction of partition values for platinum complexes, it is interesting to see if the surface area methods are robust enough to model inorganic and organic molecules simultaneously. The ability to combine organic and inorganic data would allow larger datasets to be modelled due to the more widespread availability of data for organic molecules. A wider range of molecules in the dataset will also lead to more confidence in conclusions drawn from the t-ratios, as the inorganic datasets are very limited in size. The organic data is that used to generate models of $logP_{oct}$, $logP_{CHCl}$ and $logP_{cyc}$ in chapter 4.1, (note that chlorpromazine and miconazole have been removed at they were noted to be outliers in previous surface area models).

From the results of previous studies one serious problem can be envisaged in the simultaneous modelling of organic and inorganic data, organic molecules have been shown to require scaling in order to be modelled correctly, while our scaling factors are inappropriate for inorganic molecules. In order to assess this problem three models were attempted:

i) No scaling was applied

ii) Scaling for all molecules

iii) Scaling of organic molecules only

The results for the models of $logP_{oct}$ and $logP_{CHCl}$ are given below

Table 4.14: $logP_{oct}$ and $logP_{CHCl}$ surface area models with organic and inorganic molecules

$logP_{OCT}$

	N	R^2	RMSE	f-ratio
Scaled six parameter	147	0.73	0.88	64.0
Unscaled six parameter	147	0.80	0.77	92.5
Scaled organic and unscaled Inorganic six parameter	147	0.81	0.74	100.9
Organic only model scaled	108	0.85	0.45	112.2

$logP_{CHCl}$

	N	R^2	RMSE	f-ratio
Scaled six parameter	122	0.86	0.63	117.9
Unscaled six parameter	122	0.74	0.85	55.3
Scaled organic and unscaled Inorganic six parameter	122	0.90	0.52	179.3
Organic only model scaled	108	0.92	0.42	226.8

The $logP_{oct}$ models show that no combination of scaled and unscaled descriptors explains more that 81% of the variance of the original dataset. It should be noted that the $logP_{CHCL}$ data amalgamation of the organic and Pt datasets causes the heavy correlation between the hydrogen bond acidity and basicity descriptors to be removed, consequently the splitting of PSA is possible.

For $logP_{CHCL}$ the un-scaled descriptors are the least accurate, this is not unexpected, as models constructed for organic molecules show that un-scaled descriptors were unsatisfactory and they represent 89% of the dataset. When a model is constructed using only scaled descriptors it is seen to be less accurate than a model constructed using scaled descriptors for organic molecules and un-scaled descriptors for Pt complexes, whereas the models constructed for only Pt complexes showed no preference for un-scaled or scaled descriptors. This result implies strongly that the range of functional groups within the Pt

complex dataset is so restricted that the regression is capable of fitting the error in the scaling factors for the Pt complexes in a model of only platinum complexes.

It can be concluded from these results that the original scaling factors are not suitable for modelling inorganic complexes. In an attempt to obtain more accurate scaling factors, two new fragments were added to account for Pt-NH_3 and Pt-Cl. As no experimentally observed A and B values are available for these fragments, our scaling values were obtained instead from the theoretical values calculated by Robertazzi [26]. Robertazzi calculated that cisplatin should display an A value of 0.70 a value stronger than most monofunctional acids, and a B value of 0.84, again rather stronger than most monofunctional bases. Robertazzi also noted that these properties were largely due to electrostatic effects, i.e. its hard acid/base with almost negligible covalent overlap. For these reasons we define our scaling factors thus: H in Pt-N-H is scaled by 0.12 in ASA, and Cl in Pt-Cl a scaling factor of 0.42 in BSA. While it is unusual to include Cl in our definition of hydrogen bond basicity, the values calculated by Robertazzi support its inclusion. All other polar atoms in the molecules are scaled, as before, using the principle that in larger molecules polar atoms located at a sufficient distance will mimic the nature of polar atoms in organic molecules. With these scaling factors in place the following models were constructed.

Table 4.15: $logP_{oct}$ and $logP_{CHCl}$ models created with the application of inorganic scaling factors

	N	R^2	RMSE	f-ratio
$logP_{oct}$ six parameter new scale	147	0.754	0.845	71.852
$logP_{CHCl}$ six parameter new scale	122	0.93	0.454	241.32

The $logP_{oct}$ model shows mild improvement to that of the original scaling method but is still inferior to the un-scaled models. The $logP_{CHCl}$ model yields a marked improvement of 7% in R^2 when compared to the original six-parameter model. With the resultant model now being superior to that of the model created using the combination of scaled and unscaled descriptors.

The new scaling factors are seen to be successful in the chloroform dataset and yet fail for the octanol model this difference can be attributed to the fact that the chloroform dataset

contains predominantly complexes where the only polar atoms attached to the Pt core are covered within our limited fragment definitions. Many of the molecules contained within the octanol dataset contain polar atoms bonded directly to the core that are not categorised within our definitions. As no further published A and B data is available for these fragments, our solution was to restrict the dataset to contain only Pt complexes where either N or Cl were bonded to the Pt core. With the removal of all unsuitable molecules a dataset of 18 Pt complexes remained. These 18 molecules were again combined with the dataset of 108 organic molecules. This dataset was modelled and the results are given in table 4.16.

Table 4.16: surface area models of $logP_{oct}$ dataset with restricted number of platinum complexes

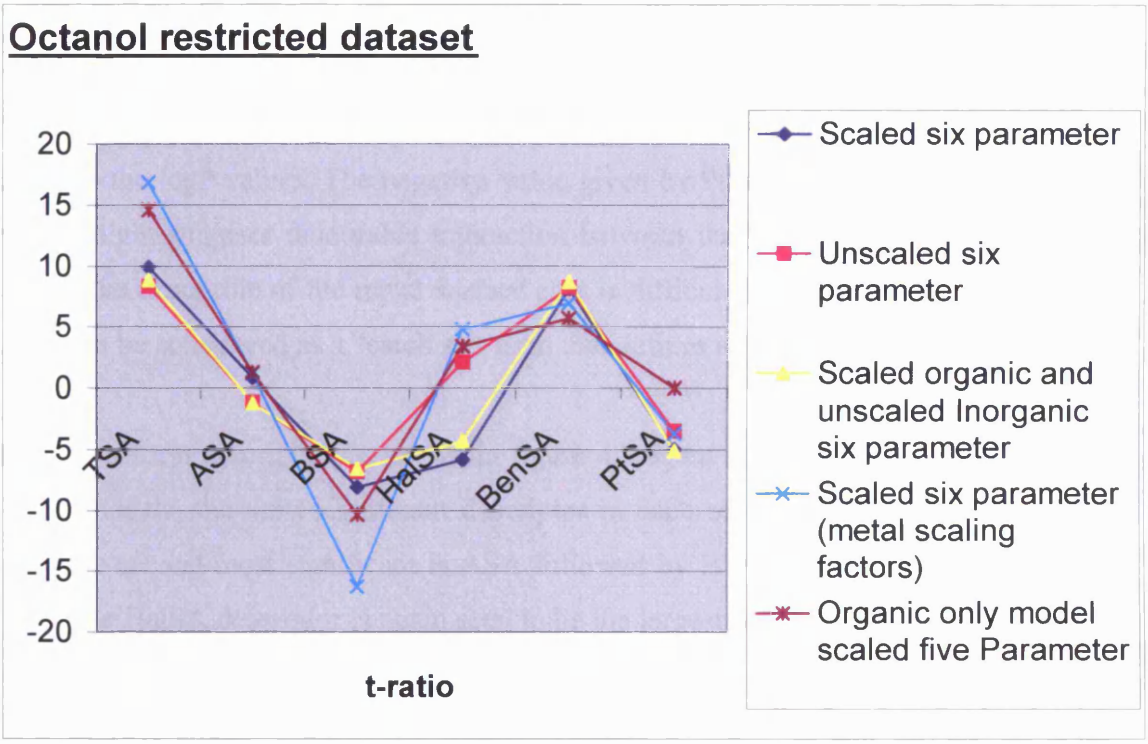
	N	R^2	RMSE	f-ratio
Scaled six parameter	126	0.804	0.645	81.687
Un-scaled six parameter	126	0.806	0.641	82.844
Scaled organic and un-scaled Inorganic six parameter	126	0.783	0.680	71.792
New scaling factors 6 parameter	126	0.906	0.446	192.872
Organic only model scaled	108	0.846	0.450	112.241

The results clearly show that the best model is that in which the new scaling factors are applied, with the model giving higher R^2 and lower RMSE values than any other combination of descriptors, the model also gives a higher R^2 and lower RMSE value than the dataset of organic only models.

4.10.6 Coefficient Analysis

Via term-by-term analysis of the t-ratios it is possible to determine if the coefficient values of the descriptors in our new models accurately represent the known chemical properties of the water/octanol and chloroform/octanol partitions. The graphs below show the plots of t-ratios for all descriptors in different models.

Figure 4.7: t-ratio values of a selection of $logP_{oct}$ models



The graph shows that the descriptors display broadly similar properties in all models. TSA is seen to be the largest and most positive of all coefficients, the next most significant descriptor is the BenSA with only two models being the exception, these exceptions are where the new metal fragment scaling factors are applied and where only organic molecules are used. In these models the second most influential descriptor is seen to be BSA acting to reduce logP values. The increased relative significance of the BSA descriptor in the models where the new scaling factors are applied and where the organic only molecules are considered is a more physically realistic representation of octanol water partition.

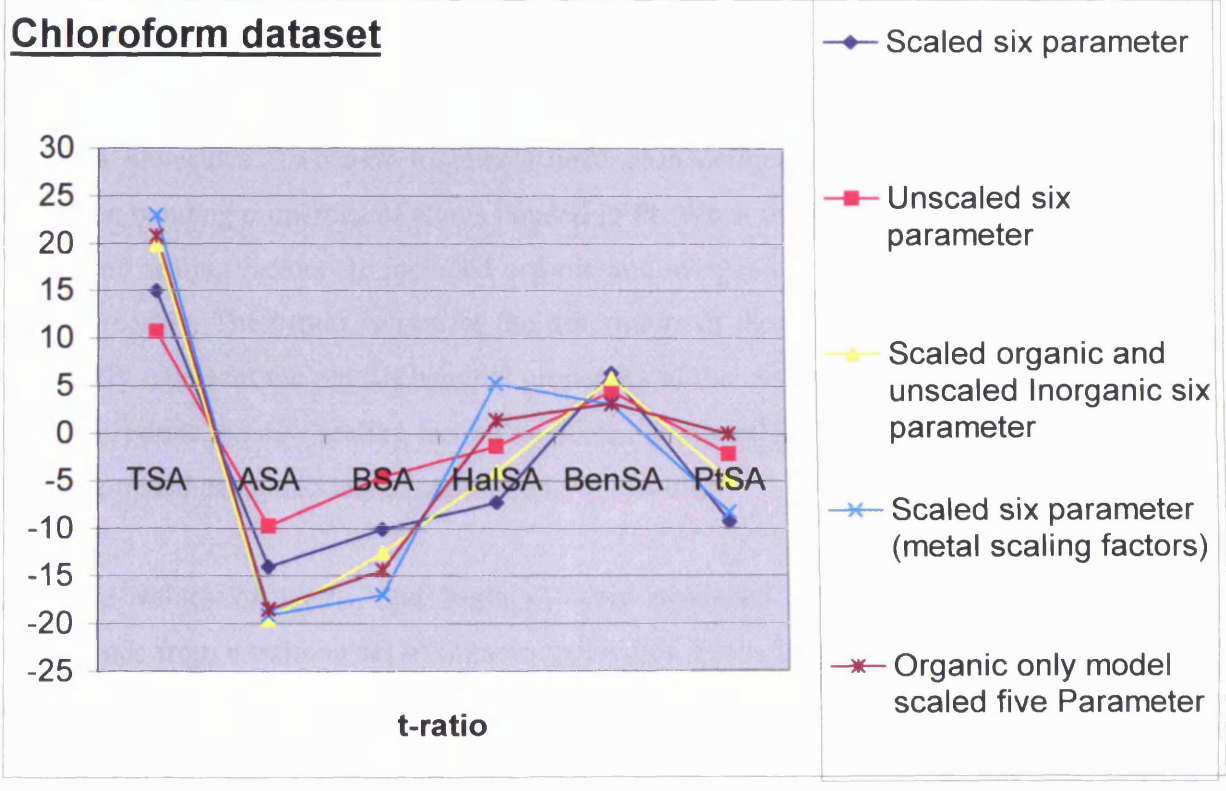
The hydrogen bond acidity descriptor is seen to have very little influence in any of the models; this result is in concordance with other models of $logP_{oct}$ and can be attributed to the similarity in hydrogen bond basicity of water and octanol. The HalSA descriptor, while of little significance, is interesting as it value fluctuates more than any other descriptor over all the models with positive and negative values being reported. This fluctuation is due to the unique hydrogen bond basicity properties of Cl atoms in cisplatin, positive values of HalSA are reported for the organic only model and the model in which

new scaling factors are used, the similarity of the two models adds further validation for our new scaling factors.

The PtSA descriptor is seen to have a small negative value throughout, thereby acting to decrease the logP values. The negative value given by PtSA in both $logP_{oct}$ and $logP_{CHCl}$ model might suggests favourable interaction between the electron poor metal atom and water. The exact role of the metal surface area is difficult to interpret without ambiguity, but it can be considered as a "catch all" term that defines all metal solvent interactions.

The chloroform descriptors as shown in figure 4.8 again show broadly similar properties for all models, the most significant descriptor in each model is TSA acting to increase logP, the second most significant is ASA followed by BSA both of which act to reduce logP. The HalSA descriptor is again seen to be the largest source of fluctuation within the model: again this can be attributed to the hydrogen bond basicity of the Cl in cisplatin. As with the octanol partition models, the positive HalSA values are reported for the organic only model and the model in which the new scaling factors are applied.

Figure 4.8: t-ratio values of a selection of $logP_{CHCl}$ models



As a final test of the new scaling factors and an overall measure of the validity of the cisplatin descriptors, the organic dataset was remodelled as a training set, the equation generated was then used to predict the logP value of the inorganic complexes. For each of these models the average absolute error in prediction of the cisplatin complexes was calculated and the results are given in table 4.17. It should be noted that for these models PtSA is not included as a descriptor as it is impossible to generate a value for it from the organic only training set.

Table 4.17: Average absolute error in prediction for platinum complex test sets

	$logP_{OCT}$	$logP_{CHCL}$
Unscaled descriptors	3.00	1.67
Scaled descriptors	4.03	4.77
Scaled descriptors (cisplatin scaled)	0.62	1.05

The results in the table show that the lowest errors in prediction are given where the metal scaling fragments are used and the worst errors are seen where the old scaling factors are applied.

4.10.7 Platinum Complex Conclusions

The results show that our surface area methods are capable of predicting logP values for cisplatin molecules. Two new fragments have been defined to help model the unique hydrogen bonding properties of atoms bonded to Pt. When these new fragments and their associated scaling factors are included organic and inorganic molecules can be modelled simultaneously. The t-ratio values for the descriptors of these models have been seen to accurately represent the physiochemical properties of the systems. The t-ratios of models that incorporate the new scaling factors most closely resemble equations from established models created using datasets of organic only molecules.

Accurate values of $logP_{oct}$ and $logP_{CHCl}$ were predicted for a test set of cisplatin compounds from a training set of organic molecules. From these results we can conclude that our scaled surface area methods are potentially capable of predicting other partitions and biological properties for cisplatin molecules with the same accuracy as organic molecules. Although this potential cannot yet be realised as the current amount of

biological partition data available for metal complexes is too limited to construct a reliable dataset.

4.11 Future Work

The identification of other important inorganic fragments and the assignment of scaling factors for these fragments would expand the number of inorganic molecules that could be treated using scaled PSA methods.

Experimentally observed partition values for cisplatin and carboplatin such as blood brain barrier perfusion and cell uptake, could be used to create models that may be highly beneficial to medicinal chemist. While this information is currently limited for inorganic molecules, the ability to combine inorganic data with the more abundant organic data could produce a sufficiently large dataset to create an accurate/robust model.

Further models of inorganic molecules could also be beneficial in determining the exact properties that are being modelled by the PtSA descriptor.

4.12 References

1. A.M. Zissimos, M.H. Abraham, M.C. Barker, K.J. Box, K.Y.J. Tam, *J. Chem. Soc. Perkin Trans. 2*, **2002**. 2. 187.
2. A.J. Leo, *MedChem software, Med. Chem. BioBytecorp.*, **2002**.
3. D.E. Leahy, J.J. Morris, P.J. Taylor, A.R. Wait, *J. Chem. Soc.-Perkin Trans. 2*, **1992**. 723.
4. J. Gasteiger, J. Sadowski, J. Schuur, P. Selzer, L. Steinhauer, V. Steinhauer, *J. Chem. Inf. Comput. Sci.*, **1996**. 36. 1030.
5. HyperChem, 6., *Published by Hypercube.inc.* 2000.
6. J.A. Platts, D. Butina, M.H. Abraham, A. Hersey, *J. Chem. Inf. Comput. Sci.*, **1999**. 39. 835.
7. JMP, *Published by SAS software.* 2000.
8. J.A. Platts, M.H. Abraham, D. Butina, A. Hersey, *J. Chem. Inf. Comput. Sci.*, **2000**. 40. 71.
9. S. Winiwarter, F. Ax, H. Lennernas, A. Hallberg, C. Pettersson, A. Karlen, *J. Mol. Graph.*, **2003**. 21. 273.
10. M.H. Abraham, F. Martins, R.C. Mitchell, C.J. Slater, *J. Pharm. Sci.*, **1999**. 88. 241.
11. www.webelements.com.
12. P. Ertl, B. Rohde, P. Selzer, *J. Med. Chem.*, **2000**. 43. 3714.
13. B. Rosenberg, L.V. Camp, J.E. Trosko, V.H. Mansour, *Nature*, **1969**. 222. 386.
14. M.S. Robillard, M. Galanski, W. Zimmermann, B.K. Keppler, J. Reedjik, *J. Inorg. Biochem.*, **2002**. 88. 254.
15. Y.A. Lee, S.S. Lee, K.M. Kim, O.L. Chong, Y.S. Sohn, *J. Med. Chem*, **2000**. 43. 1409.
16. M.J. McKeage, S.J. Berners-Price, P. Galettis, R.J. Bowen, W. Brouwer, L. Ding, L. Zhuang, B.C. Bagukey, *Cancer. Chemother. Pharmacol.*, **2000**. 46. 343.
17. D. Screnci, M.J. Mckeage, P. Galettis, T.W. Hambley, B.D. Palmer, B.C. Baguley, *Br. J. Cancer*, **2000**. 82. 966.
18. L.R. Kelland, S.Y. Sharp, C.F. O'Niel, F.I. Raynaud, P.J. Beale, I. Judson, *J. Inorg. Biochem.*, **1999**. 77. 111.
19. E. Wong, C.M. Giandomenico, *Chem. Rev.*, **1999**. 99. 2451.
20. S.Y. Loh, P. Mistry, L.R. Kelland, G. Able, K.R. Harrap, *Br. J. Cancer*, **1992**. 66. 1109.
21. R.A. Conradi, P.S. Burton, R.T. Borchardt, *Lipophilicity in Drug action and Toxicology*, ed. V.Pliska, B. Testa, and H. Vanderwaterbeemd. 1996: Weinheim. 233.
22. J. Bowedry, *MSc.Final Year Project Cardiff University*, **2002**.
23. J.A. Platts, D.E. Hibbs, T.W. Hambley, M.D. Hall, *J. Med. Chem.*, **2001**. 472.
24. R. Wysokinski, D. Michalska, *J. Comput. Chem.*, **2001**. 22. 901.
25. P.N.V. Pavakumar, P. Seetharamulu, S. Yao, J.D. Saxe, D.G. Reddy, F.H. Hausheer, *J. Comput. Chem.*, **1999**. 20. 365.
26. A. Robertazzi, J.A. Platts, *J. Comput. Chem.*, **2004**. 25. 1060.

Chapter 5 Biological Applications

5.1 The uptake of Volatile organic compounds into the cuticular matrix of plants

5.1.1 Introduction

The uptake of volatile organic compounds (VOC) from air in to plant foliage is an area of great concern; the absorption of unwanted chemicals into vegetation provides a potential path into the human food chain. This problem is exacerbated by the increasing volume of VOC that are being released into the atmosphere by industrial processes [1]. Further insight into the mechanism of uptake of VOC and the ability to envisage potentially dangerous molecules via predictive modelling would offer greater accuracy in risk assessment and help avert potential environmental disasters. There have been a number of studies and models proposed for this partition.[1-7]

Schönherr and Riederer[8] proposed that the absorption of VOC into plants is determined so predominantly by the partition of the molecule from the air into the plant's cuticle, that it is the only obstruction that must be considered. The cuticle is a protective layer that regulates and controls the exchange of water and gases from the plant, the cuticle can also help prevent certain disease causing organisms from infecting the plant but the primary role of the cuticle is to prevent dehydration.

Riederer *et al*[8] used chloroform to dewax the membranes by extraction in order obtain only the cuticular polymer matrix membrane (MX). Measurements of partition from air into isolated cuticular matrix membranes were then obtained for 50 VOC by headspace analysis; this partition is denoted as K_{MXa}. A further partition between water-solvated VOC deposited upon the cuticle and the cuticular matrix K_{MXw} was also calculated by the relationship.

$$K_{MXw} = K_{MXa} - K_{aw} \quad (5.1)$$

Where K_{aw} is the gas water partition coefficient, K_{aw} values were either determined[8] or available[5] for 40 of the 50 VOC.

The plant cuticular matrix is made of a waxy substance called cutin, which has been seen to be lipophilic in nature. For this reason it has been proposed that $\log P_{oct}$ should act as a suitable model for the cuticual matrix. Riederer *et al.*[8] showed a linear correlation of 0.819 (R^2) between $\log K_{MXa}$ and air octanol partition ($\log K_{oa}$) for 38 VOC. This correlation shows that while there is similarity between the octanol phase and the plant cuticular matrix, octanol is not similar enough in properties to act as a fully comprehensive model. Riederer *et al* also proposed a more detailed model of log K_{MXa} based on the following descriptors, molar refractivity and lypophillic contributions to water-octanol partition, a topological index descriptor and a hydrogen bond acidity term. A model based on 49 log K_{MXa} produced using these descriptors gave an R^2 value of 0.812 and a standard deviation of 0.271. While the statistics are comparable to those given by the simple plot of log K_{oa} against log K_{MXa} a strength of the model was that the physical properties that govern the mechanism of air plant cuticual matrix partition could be determined by analysis of the descriptors coefficient values. From this analysis it was seen that the plant matrix displayed hydrogen bond basic properties with hydrogen bond acidic molecules being more readily absorbed.

Further log K_{MXw} and values of log K_{Cw} (water cutical partition) have been reported by Sabljic *et al*[2] for 15 molecules in four different plant species. An interesting result of this study was that values of log K_{MXw} and log K_{Cw} were almost identical. The study also showed that values of log K_{MXw} and log K_{Cw} were fairly consistent across all four plant species. This result is in opposition with the work of Keymeulen *et al* [6] who showed that there was marked difference in uptake of VOC into cuticles of different plant species (*Hedera helix* and *Buxus sempervirens*).

Platts *et al*[9] offered a model of log K_{MXw} and log K_{MXa} for 62 molecules, using data obtained from Reiedrer *et al*[8] and Sabljic *et al.*[2] The relevant K_{aw} values needed to calculate K_{MXw} from K_{MXa} were predicted using the Abraham LFER.[10] Descriptors for this

model were not experimentally derived but calculated via the group contribution of Platts.[11] The study produced the following equations.

Air-plant cuticle partition

$\log K_{MXa} = -0.641 + 1.310S + 3.116A + 0.793B + 0.877L$ (5.2)

$N= 62\ R^2 = 0.994\ R^2_{cv} = 0.992$, RMSE = 0.230 F = 2361

Water-plant Cuticle partition

$\log K_{MXw} = -0.415 + 0.596E - 0.413S - 0.508A - 4.096B + 3.908V$ (5.3)

$N= 62\ R^2 = 0.981\ R^2_{cv} = 0.972$, RMSE = 0.236 F = 566

Both of these equations show excellent correlation between experimentally observed calculated partition values. Term by term analysis of the coefficients for equation 5.2 and 5.3 revealed a large amount of information about the partition process.

It was seen for Air plant cuticle partition (equation 5.2) that the E descriptor was insignificant and that the dispersion effects of π and n electrons had little effect on partition. The most significant descriptor within the processes was A, the hydrogen bond acidity descriptor, indicating that a molecule with strong hydrogen bond acidity would be preferentially absorbed into the cuticular matrix, and that the cuticular matrix displayed hydrogen bond basic properties, a result in agreement with the equation stated by Riederer *et al.*[8] From the small positive value of the B descriptor Platts *et al*[11] stated that the cuticle must also display some hydrogen bond acidity. The relatively large value of S reveals that the cuticle is polar/polarisable in nature. A positive value was given by the size term L suggesting that larger molecules are more preferentially absorbed e.g. via dispersion.

Similar analysis was performed for water-plant cuticle partition (equation 5.3). Equation 5.2 describes only the plant cuticle where as equation 5.3 gives the relative difference between the plant cuticle and bulk water. The coefficients revealed that cuticular matrix interacts more through π and n electron pairs and the cuticle is also less hydrogen bond basic and considerably less hydrogen bond acidic than water. The positive value given by the size term V indicates that cavities are much more easily formed in the cuticular matrix than water.

In order to explore the potential of the scaled PSA descriptors in modelling biological properties, models were created for the both datasets used in the study Platts *et al.*[9] Due to the high quality of the models produced by Platts *et al* (R^2 = 0.981 and 0.994) it is not expected that any statistical improvement will be seen, instead the models are intended to act as a test bed to assess if scaled PSA models can be created for biological data, and if the physiochemical information obtained from them is reliable.

5.1.2 Methods

Scaled PSA descriptors were generated for the 62 molecules used by Platts *et al* (1) and models were generated via MLRA. The descriptors were generated using the method outlined in 4.9.

5.1.3 Results and Discussion

The regression of $\log K_{MXa}$ and $\log K_{MXw}$ against scaled surface area descriptors produced the following equation.

$$\log K_{MXa} = -0.798 + 0.016TSA + 0.277ASA_S + 0.031BSA_S + 0.014HalSA + 0.045BenSA \quad (5.4)$$

N= 62, R^2= 0.941, RMSE = 0.729, R^2_{cv}= 0.916, F=177.982

$$\log K_{MXw} = -1.842 + 0.022TSA - 0.095ASA_S - 0.028BSA_S + 0.010HalSA + 0.0136BenSA \quad (5.5)$$

N= 62, R^2= 0.898, RMSE = 0.541, R^2_{cv}= 0.842, F=98.1977

While the overall statistics for these model are good they are slightly inferior to those proposed by Platts *et al.*[9] Partition values were predicted accurately for all molecules in both datasets with the exception of 4-nitrophenol for the air/cuticular matrix partition, where a residual of 2.369 log units was reported. 4-nitrophenol is predicted with a residual of 0.605 log units by the Abraham descriptors using equation 5.3, as this residual is so low we can assume that the experimentally observed data is reliable. As the error in prediction dose not lie in the experimental value it could be deduced that the error is generated within the scaled surface area descriptors, although the scaled surface area descriptors have been proven to be capable of predicting $\log P_{oct}$, $\log P_{CHCl}$ and $\log P_{cyc}$ values accurately for 4-Nitrophenol (see chapter 4.2). As the precise reason for the high

residual value cannot be determined the removal of 4-nitrophenol from the dataset cannot be justified.

The t-ratios of equation 5.3 and 5.4 are presented in Table 5.1.

Table 5.1: t-ratios of equations 5.3 and 5.4

Descriptor	$\log K_{MXa}$ t-ratio	$\log K_{MXw}$ t-ratio
Intercept	-2.352	-7.315
TSA	6.341	11.776
ASA_S	7.228	-3.347
BSA_S	6.554	-7.881
HalSA	6.993	6.921
BenSA	11.700	4.688

The t-ratios of table 5.1 are similar to those of the coefficient values of equations 5.2 and 5.3. For the air/cuticle matrix partition a large ASA_S t-ratio is seen, indicating that the cuticle has hydrogen bond basic properties. The value of BSA_S is higher than that reported by Platts *et al* but the value is still positive indicating that the cuticle displays some hydrogen bond acidic properties. The positive size descriptor TSA also indicates that larger molecules will show a preference to partition into the cuticle.

The most significant of all descriptors is BenSA, The exact nature of this descriptor is complex. The cuticular matrix is comprised of predominantly (40 to 80% by weight) of cutin, a high molecular weigh polymer of C16 and C18 hydroxalkanoic acids.[12] Hydroxyalkanoic acids which would not be expected to interact with benzene via π-π interactions strongly enough to merit the high significance of this descriptor. However, a study of polycyclic aromatic hydrocarbons[13] (PAH) noted that PAH with few carbon rings exhibited lower lipophilicity, higher vapour pressures and lower affinity for particle adsorption onto the cuticle surface than PAH with numerous aromatic rings. It can be assumed that the BenSA descriptor is describing a composite of all these factors. If one compares the molecules limonene and perylene, perylene displays a substantially higher $\log K_{MXa}$ value (8.35 log units higher). Yet with the exception of BenSA the other descriptors are fairly similar, TSA values of 230.54 and 291.06 $Å^2$ are calculated for

limonene and perylene respectively. Neither molecule has any hydrogen bond acid/base properties or halogen atoms. The large difference in partition is accounted for entirely by the BenSA descriptor that is calculated to be 155.42 ÅÅ^2 for Perylene and 0 for limonene.

The descriptors for logK_{MXw} are of particular interest as they reveal the relative difference between the cuticular matrix and bulk water. The plant cuticle matrix is seen to be substantially less acidic than bulk water and slightly less basic. The value of the TSA t-ratio shows that larger molecules will preferentially be absorbed into the cuticle matrix indicating that cavities are more easily formed within the cuticle matrix than bulk water. TSA is also determined to be the most significant factor determining partition. The BenSA descriptor is much less significant than in the model of air/cuticle matrix partition possibly due to factors such as gaseous adsorption onto the cuticle surface being irrelevant.

5.2 Partition into Biological liquids and tissues of vapours and biological liquids

5.2.1 Introduction

When designing a new pharmaceutical product more factors than just the molecules potency towards the biological target must be considered. The molecules ADME (absorption, distribution, metabolism and excretion) properties and toxicity must be considered if a viable product is to be developed. Of the many different biological partitions and processes that must be considered, two of the most detrimental in drug design have been determined as the partition between the blood stream and the brain/central nervous system and the absorption of molecules via the intestine. A large quantity of research has been undertaken in these processes and many models and methods of theoretical prediction have been proposed; both of these partitions are discussed in further detail in 5.3 and 5.4. The importance of other biological properties such as solubility in different biological fluids and tissues must not be overlooked.

While the ability to predict properties of drug molecules from structure and eliminate unsuitable candidates early in the design process is a very useful tool the importance of the quantitative knowledge that can be obtained from QSAR must not be ignored, and information about solubility of gases into biological phases is particularly valuable within fields such as anaesthesiology, pulmonary and hyper baric physiology. In order to obtain a large array of information about different biological environments scaled surface area models have been developed for the solubility of vapours and gases in the following liquids and tissues: blood, plasma, brain, muscle, lung, liver, kidney and heart.

There has been a large quantity of work already performed and several reviews [14-16] within which biological solubility of gases are discussed and models have been proposed. The vast majority of these correlations have been between Ostwald solubilities in biological media against Ostwald solubilities in water, fat and olive oil. [14-18] One such correlation is that offered by Sato *et al*[18] in which the solubility in blood ($\log L_{Blood}$) was correlated against the log of the product of the solubility in water and the solubility in oil. Sato *et al* derived two equations, one for the prediction of chlorinated hydrocarbons based

on 20 molecules and one for aromatic hydrocarbons based on 10 molecules. Both models were seen to be accurate with R^2 values of 0.935 and 0.861, while the equations stated by Sato *et al* show good correlations the equations are very specific to molecule types and not robust enough to act as a generic model.

Abraham *et al* [19] proposed the following model of log*L* for biological tissues again using the log*L* values of water and oil.

$$\text{Log}L_{\text{fluid/tissue}} = c + w \log L_{\text{water}} + o \log L_{\text{oil}} \qquad (5.6)$$

Where the coefficients *c, w* and *o* are regression coefficients. The equation was applied to a wide range of biological fluids and tissues including blood, plasma, brain, muscle, lung, liver and kidney. The models produced were of a very high quality with R^2 values ranging from 0.974 to 0.990 for solubility into muscle tissue and blood respectively. While these models are predictively accurate they offer very little insight into the mechanisms governing partition, the only information that could be attained was if the phase was more hydrophilic or lipohilic in nature. The models also require experimentally observed water and oil solubility values. A later study by Abraham *et al*[20] applied the Abraham LFER to the same datasets as used in their earlier study. For these models the $\log L_{16}$ (gas hexadecane partition coefficient) descriptor was used in place of the size term V. Excellent models were produced with high R^2 values and low standard deviations. From this work a comprehensive amount of information was obtained about the dynamics of the partition of gases. Abraham showed that plasma was similar to water although a little more lipophilic and that blood, lung, kidney, muscle, and brain in that order become more lipophilic less dipolar/polarisable, less acidic and less basic. Abraham also produced an LFER in which McGowan's characteristic volume was used to analyse partitions between phases such as water/phase and blood/phase.

All of the models discussed thus far have relied in some form on experimentally observed data as a descriptor. For this reason our proposed scaled surface area model offers a unique advantage in that no experimentally observed data is required allowing faster prediction of properties without need for synthesis.

5.2.2 Method

Scaled surface area descriptors were generated for the 11 datasets used in the studies of Abraham[19,20] and were created using the methods stated in 4.9.

5.2.3 Results and Discussion

In the LFER study of Abraham *et al*[20] the data obtained from the studies of Pezzango *et al*[21] and Perbellini *et al*[22] was noted as being potentially erroneous as log$L$ values reported for alkanes and cycloalkanes in blood were seen to be higher in the work of Pezzango that those of Perbellini. Via construction of a model of logL_{blood} created without values of Pezzango *et al* or Perbellini, Abraham *et al* predicted values for alkanes and cyclohexanes, the values of Perbellini were seen to be more preferable to the LFER method so with reluctance Abraham excluded the values of Pezzango from all models.

In contrast in the model of solubility of liquid non electrolytes in blood produced by Kamlet *et al*[23] no discrimination was seen against the results of Pezzango *et al*. The importance of data credibility cannot be understated, especially when one considers that Abraham and Kamlet reached differing conclusions about the nature of blood, more specifically the role of haemoglobin, based on the difference of datasets. Also the removal of all Pezzango's data represents a hefty reduction in size for some of the smaller datasets most notably that of logL_{Heart} in which Pezzango's data accounts for almost a third of the data.

In order to assess the reliability of the data of Pezzango for ourselves, two sets of models were constructed, only the first set of models contained the values of Pezzango. Table 5.2 contains the statistical analysis for a selection of these models. The results showed overall that models were very similar with or without the values of Pezzango. Analysis of the residuals shows that our scaled surface area method predicts the experimentally stipulated values proposed by Pezzango accurately with an average error of -0.209 for alkanes and 0.315 for cyclohexanes. For this reason we have chosen not to omit the values of Pezzango from our analysis. By retaining this data the distribution of the descriptors and log$L$ values covered by our models is wider than those of Abraham.

Table 5.2: Surface area models to assess the quality of the data of Pezzango

	Including the values of Pezzango			Excluding the values of Pezzango		
Partition	n	R^2	RMSE	n	R^2	RMSE
Brain	44	0.71	0.70	35	0.71	0.78
Muscle	45	0.74	0.69	36	0.75	0.76
Heart	25	0.64	0.47	16	0.67	0.58
Kidney	39	0.67	0.68	30	0.68	0.76

Abraham *et al* [20]stated that the inorganic gases display distinctly different properties in water than organics[24], and hence we have removed all inorganic gases from the models.

Three molecules were seen to be large outliers in several different systems; these were propanone, butanone and sulphur hexaflouride. These molecules were seen to give consistently poor calculated values over all models. It can be proposed that due to the inorganic nature of sulphur hexaflouride more accuracy could be obtained if specific scaling factor were added, but very little information is available to generate scaling factors, also the inclusion of such a specific fragment would begin to remove generality from our method. For these reasons sulphur hexaflouride was removed from all models.

Propanone and butanone were seen to give average residuals of 2.01 and 1.44 log units respectively. Previous models of $\log P_{oct}$, $\log P_{CHCl}$ and $\log P_{cyc}$ (see 4.2) have shown the scaled surface area descriptors capable of modelling these molecules with a maximum error of about half a log unit. It is also notable that other aliphatic ketones such as pentanone, hexanone and heptanone are predicted with high accuracy for the three datasets in which they occur ($\log L_{blood}$, $\log L_{water}$ and oil). This evidence suggests some of the values of propanone and butanone are inaccurate. For these reasons propanone and butanone were omitted from all regressions.

With the removal of all the outliers the regressions were repeated, the statistical results for all of the models produced are given in table 5.3 the modes were produced with the descriptors TSA ASA_S BSA_S HalSA and BenSA.

Table 5.3: Statistics results for all the models

	N	R^2	RMSE	F
$\log L_{Water}$ *1	81	0.649	0.626	14.105
$\log L_{Blood}$	35	0.847	0.54	32.223
$\log L_{Plasma}$	32	0.848	0.64	37.78
$\log L_{Brain}$	41	0.832	0.5	34.6
$\mathrm{Log}L_{Muscle}$	42	0.865	0.47	46
$\log L_{Lung}$	35	0.847	0.54	32.22
$\log L_{liver}$	32	0.61	0.43	13.56
$\log L_{Kidney}$	37	0.81	0.497	34.175
$\log L_{Heart}$	25	0.642	0.468	12.119
$\log L_{Fat}$	38	0.81	0.45	28.84
$\log L_{Oil}$	95	0.809	0.633	75.43

*1 – excluding all inorganic gases.

From these results it is apparent that all of our models are less accurate than those produced in the study of Abraham[20] where an average R^2 value of 0.97 and an average RMSE value of 0.19 log units was reported. While the Abraham models are superior it should be remembered that in the study of Abraham the descriptors were experimentally observed, experimental observation of descriptors requires more time and money than simple theoretical calculation making our models faster and more cost effective. While the statistics are not as good as those for non-biological phases such as $\log P_{oct}$ and $\log P_{CHCL_3}$ they are still acceptable when one considers the large experimental errors that can be associated with obtaining $\log L$ values for biological fluids and tissues.

The vast majority of models given in table 5.3 display good R^2 values of over 0.8 and RMSE values of under half a log unit. Three models are seen to give poorer fits these are $\log L_{Water}$, $\log L_{liver}$ and $\log L_{Heart}$. With the exclusion of the three aforementioned outliers none of the remaining molecules are modelled poorly enough to be classed as outliers. Stepwise regression for these three datasets reveals that the accuracy of the models can be increased significantly if the HalSA descriptor is substituted with the ClSA descriptor (total Chlorine surface area). For example substitution of ClSA for HalSA in the $\log L_{Heart}$ regression yields an increase in R^2 from 0.64 to 0.83 accompanied by a decrease in RMSE of 0.15 log units.

Further regression for all datasets reveals that Substitution of HalSA with ClSA yields an increase in R^2 for nearly all models. The models of $\log L$ in which the most improvement

is seen are those where the Abraham polar/polarisability descriptor S is seen to be highly significant, and the hydrogen bond acidity descriptor A was reported to be of little significance. As the HalSA and ClSA descriptors encompass no information about hydrogen bond acidity it can be deduced that the improvement provided by ClSA is due to better modelling of polar effects. Table 5.4 contains the experimentally observed Abraham values for a variety of halogenated organic molecules. Table 5.4 shows clearly that different halogen atoms display different polar effects. It should also be noted that the percentage of halogenated molecules within these log$L$ datasets is proportionally higher than usually found in datasets, and that the datasets also contain a wider range of halogen atom types.

Table 5.4: Varied polarity/polarisability of different Halogenated methanes

Name	S	Ref
Tetrafluoromethane	-0.25	[25]
Tertachloromethane	0.38	[26]
Tetrabromethane	0.94	[27]

For the models of logL_{liver} and logL_{Heart} a pattern is seen between halogen atom types in a molecule and predicted error, with the majority of molecules that are under predicted contain numerous fluorine atoms while a high proportion of the over predicted molecules contain several chlorine atoms.

From this evidence we can deduce that HalSA may not be the most suitable descriptor when a large portion of the molecules in the dataset contain halogen atoms of different types and where polarity is highly significant. For these reasons it was chosen to perform the regression again this time using the ClSA descriptor in place of HalSA, as before the relevant outliers were removed. The statistical results and t-ratios for these regressions are given in table 5.5. The effects of the t-ratios on different tissues and fluids is also shown in figure 5.1

Figure 5.1: The significance of descriptors within different biological tissues

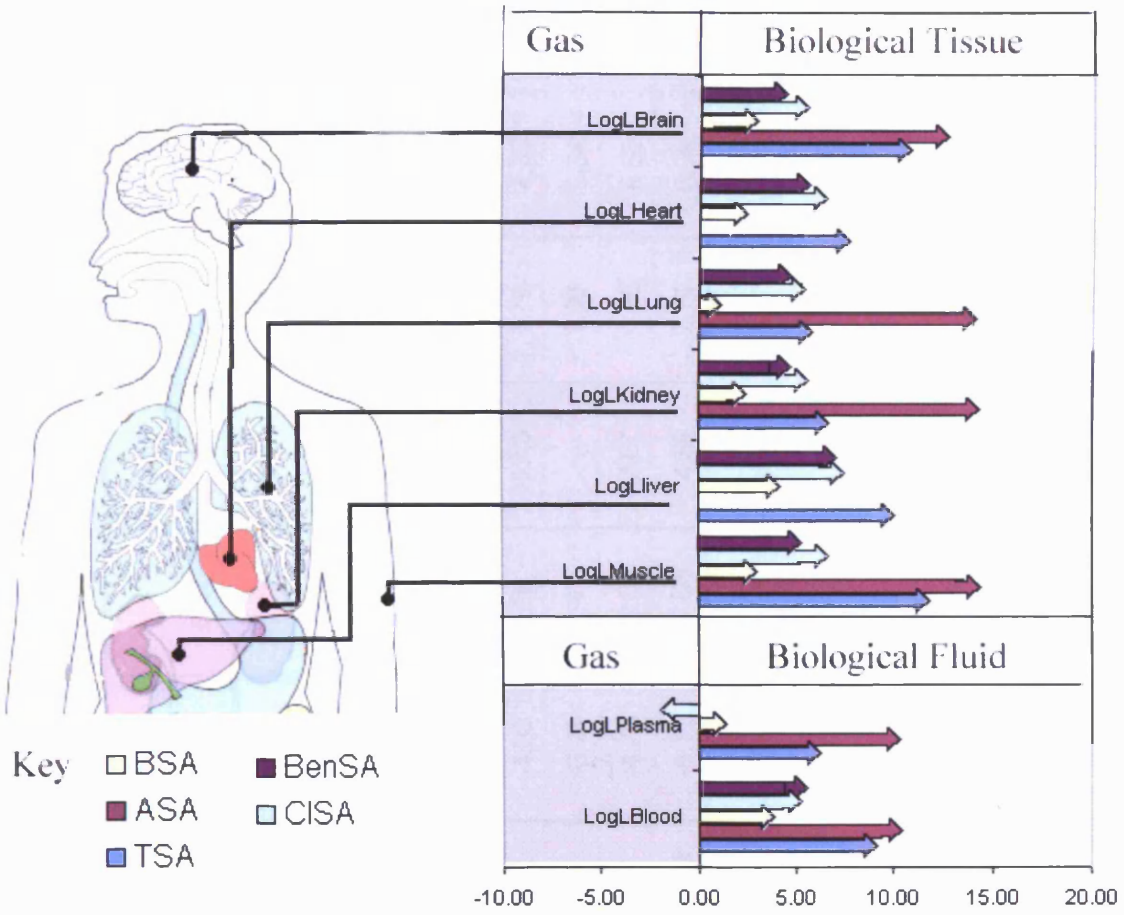


Table 5.5 Statistical analysis and t-ratio values of tissue solubility models

	R2	RMSE	N	F	Intercept	TSA	ASA	BSA	CISA	BenSA
$\log L_{Water}$ *1	0.742	0.706	81	43.14	1.11	-4.27	2.54	10.25	5.77	6.87
$\mathrm{Log}L_{Blood}$	0.72	0.66	89	43.1	-8.13	8.44	9.38	2.73	4.34	4.4
$\mathrm{Log}L_{Plasma}$	0.85	0.646	32	38.13	-6.3	5.26	9.35	0.53	-1.3	.
$\mathrm{Log}L_{Brain}$	0.897	0.392	41	61.17	-9.53	10.87	12.7	3.06	5.45	4.45
$\mathrm{Log}L_{Muscle}$	0.903	0.402	42	67.47	-9.64	10.37	13.59	1.43	5.16	4.33
$\mathrm{Log}L_{Lung}$	0.889	0.46	35	46.63	-5.7	5.55	14.03	0.71	5.18	4.14
$\mathrm{Log}L_{liver}$	0.827	0.29	32	32.19	-6.93	9.21	.	3.13	6.44	5.85
LogL_{Kidney}	0.871	0.41	37	42.04	-4.93	5.95	13.9	1.79	4.55	3.61
$\mathrm{Log}L_{Heart}$	0.827	0.32	25	23.9	-6.24	7.41	.	1.99	6.34	5.38
$\mathrm{Log}L_{Fat}$	0.896	0.339	38	55.49	-4.79	12.38	7.87	-1.65	6.58	5.11
$\mathrm{Log}L_{Oil}$	0.867	0.525	96	138.28	-9.11	17.61	4.65	2.14	8.26	7.71

N.B. Where no t-ratio values are shown indicates the dataset contained no information this descriptors e.g. no aromatic molecules in logL-plasma dataset

Via analysis of the t-ratios of each individual descriptor we see that all log$L$ partitions have a positive value for the size descriptor TSA. The only phase where a negative value is observed is the logL_{water} partition. This result is expected, as it is known larger molecules form cavities in solvents such as oil and fat more easily than water. Indicating that the vast majority of biological fluids and tissues display more lipophilic tendencies than hydrophilic with regards to molecular size. Purely using the size term TSA as a point of reference we see that of all tissues and fluids, plasma is seen to behave most like water, an expected result as plasma is comprised predominantly of water (approx 90%). Relatively low TSA values are also given for lung, kidney tissue and blood.

For the hydrogen bond acidity and basicity terms, ASA_S and BSA_S, positive t-ratio values are seen throughout with only one exception. This is an expected and chemically sensible result, as the solubility of a gas into a medium must be aided by the acidity or basicity of that medium. ASA_S is seen to be larger in the biological tissues than the biological fluids. No ASA_S t-ratios could be calculated for liver and heart tissue, as their datasets contained no hydrogen bond acidic molecules under our definitions of hydrogen bond acidity. From the values of the ASA t-ratio we can see that hydrogen bond basicity decreases in tissues in the order lung, kidney, muscle and brain.

The model of $logL_{water}$ gives the highest value of BSA_S, a result that is in agreement with Abraham [20] who stated that of all the phases studied water should be the most hydrogen bond acidic. A negative t-ratio value is given in the model of log$L_{fat.}$ As stated earlier for the solubility of gases we would expect this value to be positive by analysis of the logL_{fat} dataset we see that N_2 (a molecule where we would expect our scaling factors to fail) has a very influential effect on the value of BSA_S. If N_2 is removed the significance of BSA_S drops to the point where it becomes insignificant and a model of equal quality can be produced with only four descriptors.

From the BSA_S t-ratio values of the oil water and fat models we can state that if a phase is hydrophilic it will be more hydrogen bond acidic than a lipophilic phase. The t-ratios for BSA_S show that blood is hydrogen bond basic. A low t-ratio value is given for BSA_S in plasma, in contrast with the earlier findings from TSA that plasma is similar in properties to water. Analysis of the dataset for logL_{plasma} shows as with logL_{fat} that nitrogen and other inorganic gases such as oxygen and nitrous oxide heavily influence the coefficient value

of BSA_S. Removal of these molecules from the regression causes the value of the BSA_S descriptor to rise to 2.15 indicating that plasma is more hydrogen bond acidic than originally calculated. The removal of the inorganic gases from the $\log L_{plasma}$ regression causes very little change in the t-ratios of the other descriptors or overall statistics of the model. For the biological tissues the hydrogen bond acidity is seen to increase in the following order lung, muscle, kidney, heart, brain and liver. Although it should be noted that for all biological tissues except brain and liver tissue the t-ratio values are bellow 2 and in insignificant at the 95% level and in predictive models the BSA_S descriptor should be removed.

As stated the ClSA descriptor accounts for the polar/polarisable effects of the phase. The largest ClSA descriptors are given by the models of $\log L_{fat}$ and $\log L_{oil}$, this is not the expected result as it would be expected $\log L_{water}$ would exhibit far more polar properties than fat or oil. Although the ClSA descriptor within the individual models of $\log L_{fat}$ and $\log L_{oil}$ is much less dominant than the size term TSA unlike the model of $\log L_{water}$. The importance of the ClSA t-ratio increases in the series kidney, muscle, lung, brain, heart and liver. The t-ratio value of $\log L_{plasma}$ is −1.30 meaning that it is insignificant at the 95% level and should be removed from a predictive model.

The aromatic surface area descriptor BenSA is seen to give the largest values in the liquids oil and water, with biological tissues giving lower values. This indicates that that water and oil are more likely to interact with π electrons. The importance of the BenSA descriptor increases in the series kidney, lung, muscle, blood, brain, heart and liver.

Similarities can be seen within the order in which the size of the BenSA and ClSA descriptor increase across all of the different models. Of all biological tissues kidney tissue gives the lowest values for ClSA and BenSA, signifying that the kidney tissue is least likely to interact with halogenated or aromatic molecules, unlike brain tissue in which the largest t-ratios are seen for ClSA and BenSA.

In order to get a clearer picture of which descriptors are important within individual models, the size term TSA was scaled to a value of one for each system and all other descriptors within the model were scaled accordingly. The hydrogen bond basicity term

BSA was not particularly dominant in any of the log$L$ systems except logL_{water} in which BSA is the most significant term.

ASA is clearly the most dominant term in logL_{lung} and logL_{kidney} and to a lesser extent in logL_{brain} logL_{muscle} and logL_{plasma}. All of these models also exhibit very low BSA values indicating that they are more hydrogen bond basic than acidic than in nature. For the remaining systems TSA is seen to be the most influential descriptor with the exception of logL_{water}. The descriptors BenSA and ClSA seem to be generally more significant than BSA but less influential than TSA and ASA. Generally BenSA and ClSA are seen as the third and forth most significant terms in each regression.

5.3 Blood brain barrier

5.3.1 The blood brain barrier Introduction.

The cerebral capillaries are distinctly less permeable than other capillaries in the body; this restriction in permeability is described as the blood brain barrier (BBB). In drug design it is important to determine whether a candidate molecule is capable of penetrating the BBB. For drugs targeted at the central nervous system BBB penetration is essential, where as peripherally acting drugs aimed at other sites of action unwanted penetration of the BBB may lead to undesirable side-effects.

While the models produced in 5.2 for solubility in blood, plasma and brain tissue are informative they do not actually yield any information about the factors and mechanisms that govern selective nature of the blood brain barrier. A common measure of a molecules ability to penetrate the BBB is the logarithm of the ratio of the concentration of the molecule in the brain over the concentration of the molecule in the blood, expressed as log (C_{brain} / C_{blood}) or logBB. Experimental methods to determine logBB require animal testing and synthesis of the compound; these experiments are not only expensive but also difficult and time consuming to perform. Artificial membrane based methods[28] for studying of BBB penetration are being developed, but these methods still require synthesis of the test molecule and are not necessarily the same as logBB.

In silico prediction of logBB can serve as a powerful tool in drug design and discovery, and hence there has been a great deal of research in to predictive models of BBB[29]. While the permeation of the blood brain barrier has been seen to be complex in nature many well established conclusions have been drawn.[30-33] Highly polar molecules are seen to exhibit lower logBB values, except in cases where the molecule in question is capable of active transport. Size, ionisation, hydrogen bonding and molecular flexibility have all also been seen to be influential in penetration of the BBB.

It has also been established that PSA is detrimental to logBB. One of the earliest models of logBB in which PSA was used as a descriptor was that of Kansy and van de waterbeemd[34] who developed a QSAR based on the regression of logBB values of 20 molecules against their PSA_U. The study gave the following equation.

$$LogBB = 1.64 - 0.21PSA_U + 0.003mol\_vol \quad (5.6)$$
$N = 20,\ R^2 = 0.697,\ RMSE = 0.448$

Where mol_vol is the total molecular volume, van de waterbeemd stated that blood brain penetration at equilibrium would be decreased if PSA_U were increased. Later applications of equation 5.6 to molecules outside the dataset showed poor predictive results. Thus indicating that a 20 compound training set would be insufficient to create a highly accurate model for prediction of logBB. Kelder *et al*[35] correlated PSA against logBB for a set of 45 drug like molecules. Kelder *et al* attempted to improve the correlation via inclusion of other calculated molecular properties such as molar weight, molecular volume, logP, dipole moment etc, none of which were seen to be significant within a stepwise regression after PSA had been entered. Correlations of dynamic PSA and logBB were also performed, but the increase in accuracy of the model was so slight that the time consuming conformational search on each molecule was not justifiable. Via these correlations and the analysis of 776 orally administered CNS drugs Kelder *et al* concluded that for high penetration of the blood brain barrier a molecule should posses a PSA of less than 60 $Å^2$.

A correlation of PSA_U against logBB was also performed by Clark[36] for a dataset of 57 molecules. This regression produced the following equation.

$$logBB = 0.547 - 0.016PSA \quad (5.7)$$
$N=57,\ R^2=0.671,\ Sd= 0.455$

Clark[36] and Platts[37] both commented that a model based on equation 5.7 would be unrealistic as it would be incapable of predicting logBB for any nonpolar molecule (Although this may be of little concern for pharmaceutical applications as no potential drug is likely to exhibit a PSA of 0 $Å^2$). To account for this Clark produced equations that used molecular weight, molecular volume and non-polar surface area none of which produced significant improvement in the predictive accuracy of the model. The best model produced used calculated logP (ClogP) values as an additional descriptor. The model gave an R^2 value of 0.787 and a standard deviation of 0.354.

Many models of logBB have been produced which were not reliant on PSA as a descriptor such as that of Lombardo[38] where logBB was correlated against the linear free energy of solvation in water. Another model that was not dependant on PSA was that of Abraham *et al*[39] who constructed a LFER model using the same training set as utilised in equation 5.7. The model gave impressive statistics (R^2 = 0.900 and RMSE = 0.201 log units). In a follow up to this study Platts *et al*[37] expanded the number of molecules in the training set from 57 to 157. The LFER of Abraham was applied again along with an additional descriptor I_1, which was an indicator variable for carboxylic acids. After removal of outliers equation 5.8 was produced. For 112 of the compounds in equation 5.8 experimentally observed descriptors were available, the descriptors for the remaining compounds were calculated using the group contribution of Platts.[11]

$$\log BB = 0.021 + 0.463\,E - 0.864\,S - 0.564\,A - 0.731\,B + 0.933\,V - 0.567\,I_1 \qquad (5.8)$$

$N = 148$, $R^2 = 0.745$, RMSE = 0.343, $R^2_{cv} = 0.711$, $F = 69$

While the statistics of this model were lower than that of the original 57 data point model, the increase in chemical diversity along with the range of descriptor values covered by the model more than compensated for the loss in accuracy.

It can be concluded that PSA is highly significant to logBB but alone it is incapable of producing highly accurate models. Methods like the those of Abraham, which encompass a wider range of chemical information and are not reliant on PSA as a descriptor are capable of predicting logBB with better accuracy. As our scaled PSA method can be thought of as a "halfway house" between PSA and methods such as the LFER of Abraham we have attempted numerous models of logBB to assess our scaled PSA method as a tool for predictive modelling and a method of determining more about the nature of PSA upon blood brain barrier partition.

5.3.2 Methods

Scaled polar surface area descriptors were generated for all of the molecules used in the datasets of Kelder *et al*[35] Abraham *et al*[39] and Platts *et al.*[37] Descriptors were calculated using the method stated in 4.9.

For clarity the three datasets will be referred to as I for the dataset of 45 molecules proposed by Kelder,[35] II for the dataset of 57 values proposed by Abraham[39] and III for the dataset of 157 values proposed by Platts.[37]

5.3.3 Results

5.3.3.1 dataset I

A number of models were constructed for dataset I using a variety of descriptors, the results of which are shown in table 5.6. Addition of TSA to the regression yields very little improvement to the statistics. Splitting of PSA_U into ASA_U and BSA_U components gives a small improvement to the model with an increase in R^2 of 0.018 and a decrease in RMSE of 0.015. After TSA ASA_U and BSA_U are inserted into a stepwise regression the descriptors HalSA and BenSA are seen to be completely insignificant at the 95% level.

Table 5.6: Statistical analysis of a number of blood brain barrier models made using dataset I (45 molecules)

Descriptors	R^2	RMSE
PSA_d	0.840	0.369
PSA_U	0.791	0.423
PSAs	0.588	0.590
TSA PSA_U	0.810	0.408
TSA PSA_S	0.595	0.597
TSA ASA_U BSA_U	0.828	0.393
TSA ASA_S BSA_S	0.672	0.543

While the splitting of PSA_U in to its components yields only a small increase in accuracy the increase makes it comparable to the equations proposed by Kelder[35] in which PSA_d (dynamic PSA) was used as a descriptor

If PSA_U in a two-parameter model with TSA is substituted with its scaled counterpart a noted decrease in accuracy is observed. The R^2 value drops from 0.81 to 0.60 and the RMSE value increases from 0.408 to 0.597. A similar decrease in statistics is seen if ASA_U and BSA_U are replaced with ASA_S and BSA_U in a three-parameter model using

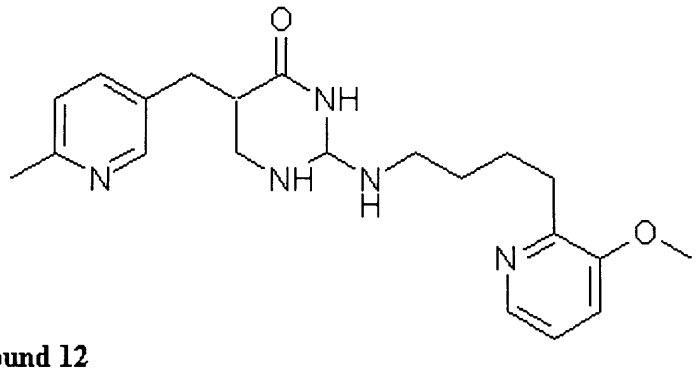
TSA. As with the previous model a stepwise regression reveals the descriptors HalSA and BenSA are insignificant at the 95% level.

5.3.3.2 dataset II

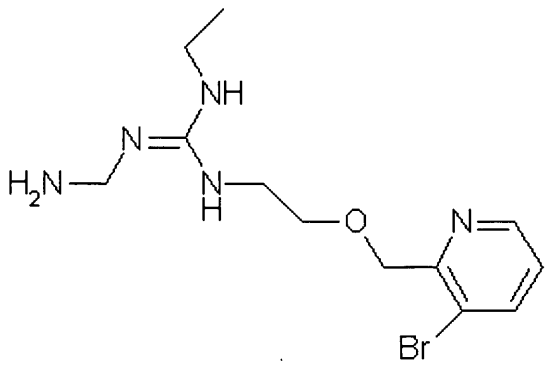
The same selection of PSA models that were produced for dataset I was created for the dataset II. The statistical analysis of these models is given in table (5.7). When Clark[36] modelled this dataset two molecules were omitted. N_2 was removed as ClogP was used as a descriptor and values could not be calculated for N_2. Compound 12 was also omitted (shown in figure 5.2) as it had been seen to be an outlier by other groups[40,41]. We have also chosen to remove these two molecules from our models as our surface area descriptors have been seen to be unreliable for diatomic gases such as N_2 (see 5.2.3), and compound 12 was also seen to be an outlier when used with our descriptors. Compound 3 (shown in figure 5.2) has also been removed as it was poorly predicted, this removal can be justified as it has been excluded from models created by other groups[39].

Figure 5.2: Structure of compounds 3 and 12

Compound 3



Compound 12



The models produced using dataset II show similar trends to those exhibited by the models created with dataset I. Splitting PSA_U into its component acid and base descriptors gives negligible improvement. When ASA_U and BSA_U descriptors are used BSA_U is seen to be the more dominant of the two. For this dataset a positive t-ratio value is given by ASA_U although this result is in conflict with our previous findings little attention should be paid, as ASA_U is so insignificant its removal, to leave a model of TSA and BSA_U gives almost identical statistics. As before stepwise regression reveals that after TSA, ASA_U and BSA_U are inserted HalSA is seen to be insignificant although now BenSA is seen to be of a small importance with its addition boosting R^2 value from 0.734 to 0.758.

Table 5.7: Statistical analysis of a number of blood brain barrier models made using dataset II (54 molecules)

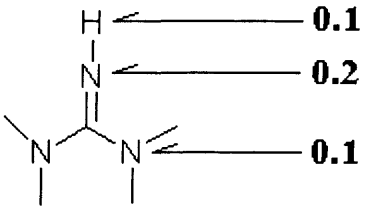
	R^2	RMSE
PSA_U ClogP	0.800	0.317
PSAu	0.706	0.389
PSAs	0.570	0.468
TSA PSA_U	0.728	0.378
TSA PSAs	0.584	0.467
TSA ASA_U BSA_U	0.734	0.378
TSA ASA_S BSA_S	0.620	0.451
TSA ASA_U BSA_U BenSA	0.758	0.369
TSA ASA_S BSA_S BenSA	0.665	0.427

The application of ASA_S and BSA_S descriptors acts only to hinder the models causing the R^2 values to drop and RMSE values to rise. BenSA is seen to have mild significance while HalSA is not important. Where ASA and BSA are scaled the significance of the two descriptors switches making ASA more the more important of the two.

Dataset II contains many molecules, which contain a guanidine functional group, a functional group that is not covered within our scaling descriptors. As the abundance of this functional group is a source of potential error within our scaling factors a fragment and scaling factors for the guanidine functional group were defined and the models were

reproduced to include this fragment. The scaling values assigned to the guanidine fragment are shown in figure 5.3.

Figure 5.3: Values of scaling factors assigned for guanidine functional group.



With the new scaling factors applied the scaled models were seen to be an improvement over the original scaled models but still giving inferior results to those where unscaled descriptors were used. For a simple three parameter model of TSA, ASA_S and BSA_S an improvement of 0.06 is seen in R^2 along with a decrease of 0.04 in RMSE. The t-ratios for these new models now show that ASA_S and BSA_S have almost identical equal effects upon blood brain barrier partition with increasing H-bond acidity and basicity acting to reduce a molecules penetration of the blood brain barrier. HalSA is seen to be insignificant and BenSA is seen to have a small role in penetration.

Results 5.3.3.3 dataset III

As with the datasets I and II a selection of models was created using a variety of descriptors. The statistical analysis of these models is displayed in table 5.8. The scaling factors used for these models incorporate the guanidine fragment defined in 5.3.3.2, as the guanidine fragment is also unusually common within this dataset. A number of molecules were determined to be outliers and removed from the models shown in table 5.8. Theses molecules were Lombardo 4, Lombardo 20, YG19, YG20, ORG12692, mesoridazine and indomethacine. The first five of these molecules have been reported and omitted as outliers in previous studies [35,36,40,42,43].

Table 5.8: Statistical analysis of a number of a variety of blood brain barrier models made using dataset III. (150 values)

	R^2	RMSE
LFER	0.745	0.343
PSAu	0.508	0.470
PSAs	0.270	0.572
TSA PSA_U	0.595	0.429
TSA PSAs	0.364	0.534
TSA ASA_U BSA_U	0.597	0.4299
TSA ASA_S BSA_S	0.41	0.51
TSA ASA_U BSA_U BenSA	0.612	0.4233
TSA ASA_S BSA_S BenSA	0.424	0.516

The models show the same pattern as those made with datasets I and II. The best models are created when the unscaled descriptors are used. Splitting of PSA_U offers no improvement to the model. When PSA_U is decoupled BSA_U is seen to be the more significant of the two descriptors. The best model produced with the unscaled descriptors is the four parameter model in which the descriptors TSA, ASA_U, BSA_U and BenSA are used, giving an R^2 value of 0.612 and RMSE value of 0.423 log units. By comparison to the model produced by Platts *et al*[37] where an R^2 value of 0.745 was reported the statistics of our model are not as disappointing as they may first appear. Platts *et al* [37]stated that due to the high chemical diversity and the range of sources from which this data was acquired models with exceptionally high predictive accuracy would not be expected to be produced.

The scaled PSA models also show the same patterns as our other scaled PSA models, with scaling of PSA decreasing the quality of the models. Slight improvement is seen when ASA_S and BSA_S are used, in place of PSA_S. For the descriptors BenSA and HalSA only the former is seen to be significant.

5.3.4 Discussion

All three datasets show that if scaling factors are applied to PSA then the quality of the models produced is reduced. The models also show that partition of PSA_U into separate acid and base descriptors offers little or no improvement to the model. While these two conclusions indicate that our scaled PSA methods are unsuitable for modelling BBB partition some conclusions can be drawn about the nature of the BBB from our failings that are informative to pharmaceutical scientists with regards the use of PSA.

The scaling factors produced in 3.1.3 have been seen to beneficial to simple chemical partitions such as $logP_{oct}$ and $logP_{CHCl}$ (4.2) and biological partitions such as water/plant cuticular matrix (5.1), but none of these partitions are as complex and selective in nature as that of the blood brain barrier.

In the model of Platts *et al*[37] an extra descriptor was added to the LFER of Abraham as an indicator of carboxylic acids. This descriptor was required as the presence of CO_2H reduced brain penetration far higher than simply its hydrogen bond and polarisability properties could account for. It was proposed that CO_2H had such low uptake by the brain because of its affinity for binding with albumin when ionised to CO_2^-, along with efflux actions within the brain acting to flush out molecules containing CO_2H. A similar indicator variable was used within the study of Salminen *et al*,[43] although this variable accounted for the presence of carboxylic and amino groups. It is more than conceivable that other functional groups have unique properties within the brain penetration mechanism. For this reason the scaling factors we have proposed to make PSA a more realistic measure of hydrogen bonding may not be applicable to such a complex partition as BBB. The generality offered by the traditional definition of PSA may be far more applicable to logBB calculation as it does not in any way attempt to compensate for specific functional groups. While PSA is established as being a representation of a molecules hydrogen bonding ability, it has been used in models of blood brain barrier in conjunction with other descriptors that encode information about hydrogen bonding, such as the study of Feher *et al* [44] where PSA was used along with logP and the number of hydrogen bond acceptors in aqueous media.

All un-scaled models show that separation of PSA into its components gives little or no improvement to the model. Analysis of the t-ratios for un-scaled three parameter models of TSA, ASA_U and BSA_U for all datasets shows that either ASA_U and BSA_U have a similar effect on blood brain barrier penetration both acting to reduce logBB and hinder perfusion into the brain in the case of dataset I, or that the BSA descriptor is considerably more significant than the ASA descriptor as in datasets II and III.

For the model made using dataset I the t-ratios show the descriptors ASA_U and BSA_U are modelling similar physiochemical effects, for this reason no significant improvement can really be expected to be gleaned from the separation of PSA_U, as the regression is capable of incorporating all the chemical properties within this one descriptor.

The models made from datasets II and III show that the ASA_U descriptor is far less significant than BSA_U. Multivariate analysis of the datasets shows that PSA_U and BSA_U correlate to more than 99%. The dominating role of BSA_U and its similarity to PSA_U within these datasets more than explain why separation PSA_U offers no improvement to the models. Feher *et al*[44] also stated that hydrogen bond basicity was more important than acidity from the observation that the number of hydrogen bond acceptors was significant within their QSAR of blood brain partition where as the number of hydrogen bond donors had little impact on the statistics of their model.

While splitting of PSA and the addition of extra descriptors offers little improvement, the best R^2 and RMSA values for the models produced for each dataset employ the descriptors TSA ASA_S BSA_S and BenSA for scaled descriptors and TSA, ASA_U, BSA_U, and BenSA for the unscaled descriptors. Although as shown in tables 5.6, 5.7 and 5.8 these models are only the best by a narrow margin of a few %. For dataset I and II a simple model of PSA_U is better than the four-parameter model as the increase in accuracy dose not merit the inclusion of three descriptors, although for dataset III the improvement of 10% in the R^2 value justifies the inclusion of extra descriptors.

Figures 5.4 and 5.5 show how the values of the t-ratios vary over the three datasets for both scaled and unscaled models, these models were chosen for analysis as the larger number of descriptors means that more information can be obtained from their interpretation. For the unscaled models broad similarities can be seen between datasets II

and III, but the t-ratios for dataset I display different properties. This is not unexpected as dataset I is a lot smaller in chemical diversity than datasets II and III. The t-ratios show that the most significant of all descriptors is either BSA_S or BSA_U always giving a negative value indicating that hydrogen bond basic atoms act to hinder perfusion into the brain. ASA_U is seen to be less significant than BSA_U and is either completely insignificant or slightly negative. When ASA and BSA are scaled ASA_S exhibits more significance and t-ratio values more closely resemble the t-ratios values of the Abraham LFER descriptors A and B as calculated from previous studies.[37] This shifting in values of the descriptors ASA_S and BSA_S to more closely resemble the A and B descriptors of Abraham upon scaling suggests again that the scaling of ASA_U and BSA_U causes them to better represent hydrogen bond acidity and basicity. However, the decrease in accuracy suggests that the predictive power of PSA_U is not due to it accurately representing a molecule's capacity for hydrogen bonding but instead acting as a more general representation of hydrogen bonding and polar effects.

The size descriptor TSA is seen to give small positive t-ratios for both the scaled and un-scaled models for dataset II and III. The positive TSA value suggests that larger molecules will be pushed out of the bloodstream and into the brain.

Figure 5.4: t-ratio values for unscaled logBB models

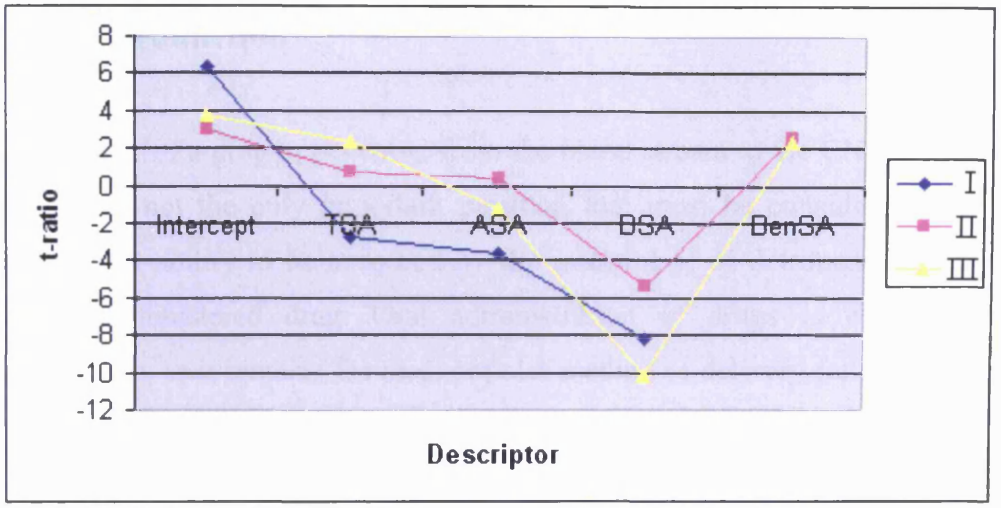
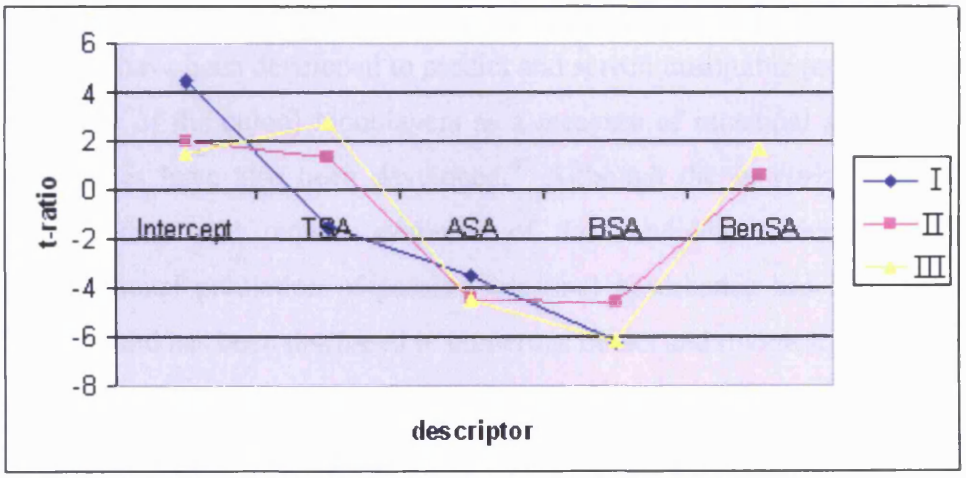


Figure 5.5: t-ratio values for scaled logBB models



Positive t-ratio values are also seen for BenSA in all models except those for dataset I where BenSA was determined to be insignificant via stepwise regression. This positive value indicates that molecules with a high degree of aromatic bonding/ π electrons will be preferentially partitioned into the brain. The previous studies in which the data from datasets II and III were modelled with the LFER of Abraham are in agreement with our findings that molecular size and the presence of π electrons will increase a molecule's absorption into the brain as both Abraham[39] and Platts [37]reported positive values for V and E in their studies.

5.4 Intestinal Absorption

5.4.1 Introduction

The ability of a drug to penetrate from the blood stream to the CNS is important, but it is obviously not the only important partition that must be considered in drug design. A molecule's ability to be absorbed by the intestine is of detrimental importance for any orally administered drug. Oral administration of drugs is of particular industrial importance, as it remains the most popular method of delivery for pharmaceuticals due to convenience, low cost and patient compliance.

Similarly to blood brain barrier penetration, experimental *in vivo* methods to determine intestinal absorbance are expensive, time consuming and often unreliable. *In vitro* methods have been developed to predict and screen unsuitable molecules, including Caco-2 (cancer of the colon) monolayers as a measure of intestinal absorbance.[45,46] Artificial membranes have also been developed.[47] Although the *in vitro* methods are faster and cheaper they still require synthesis of the candidate molecule. For these reasons computational prediction of passive intestinal absorbance has attracted a great deal of research and has been discussed in numerous books and reviews. [29,48,49]

One of the most basic models of intestinal absorption is the Lipinski's rule of five.[50] This states a molecule will be poorly absorbed by the intestine if more than two of the following chemical properties are exceeded, logP > 5, molecular weight > 500, number of hydrogen bond donors > 5 and number of hydrogen bond acceptors > 10. PSA has also been used as a measure for predicting passive intestinal absorbance. Palm *et al*[51] produced a sigmoidal fit of 20 molecules and their percentage intestinal absorption against dynamic PSA, this fit gave an R^2 value of 0.94. Analysis of the relationship revealed that a molecule exhibited poor intestinal absorbance (deemed as being < 10%) when the PSA > 140 Å^2. (see figure 1.1)

The dataset of 20 molecules used by Palm[51] was modelled again by Clark[52] although in this model static PSA was used in place of dynamic PSA. The results of this correlation were almost identical to those of Palm, concluding that static PSA was equally capable of acting as an indicator of a molecules potential intestinal absorbance.

Clark further tested the >140 Å ^{2} criterion via the scrutiny of a dataset of 74 compounds. The results concluded that PSA was indeed an accurate method of classifying a molecules intestinal absorbance and better method of classifying a molecules intestinal absorbance than lipinski's rule of five.[50]

The methods of Clark and Palm offer classification of intestinal absorption via PSA, other *in silico* methods have been developed in which values of intestinal absorbance are calculated using traditional QSAR methods. Raevsky *et al.*[53] used HYBOT descriptors to model intestinal absorbance data for 17 molecules. The model used two descriptors based on hydrogen bond acceptor and donor properties. The model used a non-linear fit and gave a $R^2 = 0.954$.

Wessel *et al*[54] used a genetic algorithm for descriptor selection along with a neural network to calculate intestinal absorbance for 74 molecules (the same dataset that was utilised in the study of Clark[52]). While a good fit of observed against calculated intestinal absorbance values was given by this model, the complexity of neural networks meant that very little information could be gained about the individual roles of the descriptors within this model.

There have also been a number of models developed that have been designed not only to be capable of calculating intestinal absorption but also be easily interpretable. One such example is the study of Abraham[55] in which % intestinal absorption was transformed into a first-order rate constant and regressed against the Abraham descriptors. From comparison of the regression co-efficient to other solvation equations it was suggested that the main process in intestinal absorption was diffusion through the stagnant mucus layer, together with transfer across the mucus membrane interface.

5.4.2. Methods

Three datasets of intestinal absorption were acquired from previous studies, a dataset of 20 intestinal absorption values from the study of Palm, a dataset of 74 intestinal absorption values used in the studies of Clark[52] and Wessel[54] and a dataset of 125 values from the study of Abraham *et al*[55]. For convenience the dataset of 20 values will be

referred to as dataset I, the dataset of 74 values will be referred to as dataset II and the set of 125 values dataset III.

In order to model the % intestinal absorbance using multiple linear regression analysis the data was transformed using the following equation. This form of transformation from percentage to arcsine values is necessary, as intestinal absorption has been seen to have a sigmoidal curve when plotted against PSA.[51]

$$AbsT = \sqrt{ArcSine\left[\frac{\%\ Absorbance}{100}\right]} \qquad (5.9)$$

Surface area descriptors were obtained using the methods out laid in 4.9.

5.4.3.1 Results dataset I

Multiple linear regressions were performed using different surface area descriptors against AbsT, the results of which are given in table 5.8. Foscarnet was removed from all regressions as the surface area descriptors are not intended or tested for the application to charged species. From the results produced it is evident that dynamic PSA and PSA_U when regressed against AbsT give similar models both fitting about 86% of the variance of the data. When a similar regression is performed with PSA_S the R^2 drops and the RMSE rises. While splitting of PSA_S and the inclusion of BenSA and HalSA descriptors in to the equation gives improvement, the statistics of the model its quality is still inferior to that of a single parameter model of PSA_U. A five-parameter model made using ASA_U and BSA_U gives and excellent model, with an R^2 value of 0.945 and RMSE value of 0.11.

While the results shown in table 5.8 give some insight into the relationships between the varying types of PSA (static dynamic and scaled) and intestinal absorbance little confidence should be placed in the findings, as dataset I is very limited in both chemical diversity and total number of molecules.

The scaled descriptors are seen to be unsuitable in modelling this dataset; this unsuitability is caused by the occurrence within the dataset of molecules such as Mannitol, Lactulose and Raffinose, all of which are capable of forming numerous

intramolecular hydrogen bonds. As these intramolecular hydrogen bonds occur upon aliphatic carbon chains they are not accounted for within the scaling factors. Due to the limited size and structural diversity of the dataset these molecules represent a substantial portion of the hydrogen bond acidic molecules of the dataset. This particular artefact of the dataset is also manifested within the t-ratios of the scaled five parameter model ASA_S is seen to be significant with a t-ratio value of -3.58 while BSA is seen to be insignificant, this is in contradiction with previous findings that have stated that both hydrogen bond acidity and basicity are important to intestinal absorbance. The dataset also contains hydrogen bond basic fragments such as azenes that are not incorporated into our scaling factors. These unclassified functional groups would have less influence on the coefficient values generated by the regression (or may be outliers) within larger and more structurally broad datasets.

Table 5.8: Models of intestinal absorption using dataset I (19 molecules)

	R^2	RMSE	F-ratio	R^2cv
PSAd	0.862	0.155	107.000	0.761
PSA_U	0.865	0.154	108.610	0.783
PSA_S	0.657	0.245	32.590	0.532
Five parameter unscaled	0.945	0.111	44.949	0.856
Five parameter scaled	0.821	0.202	11.950	0.580

5.4.3.2 Results dataset II

Regression of PSA descriptors against AbsT values of dataset II showed three molecules to be outliers; these were methotrexate, amoxicillin and cefuroxime axetil. All of these outliers are known to exhibit active transport properties. Methotrexate is absorbed by a carrier-mediated process, which is responsible for foliate absorption [56,57], Amoxicillin is absorbed via dipeptide carriers [58,59], and cefuroxime axetil has been noted to absorb by a specialized transport mechanism that obeys michaelis-menten kinetics.[60] As the aim of our model is to calculate only passive diffusion these molecules were justifiably removed.

The remaining dataset of 71 compounds was modelled using a selection of descriptors. The statistical analysis of the equations produced is given in table 5.9.

Table 5.9: Results of surface area descriptors models of intestinal absorption for dataset II (71 molecules)

	R^2	RMSE	F-ratio	R^2cv
PSA_U	0.43	0.25	53.509	0.39
PSA_U TSA	0.423	0.25	24.94	0.36
TSA ASA_U BSA_U	0.464	0.24	19.32	0.38
TSA PSA_U HalSA BenSA	0.443	0.25	13.11	0.36
TSA ASA_U BSA_U HalSA BenSA	0.473	0.24	11.68	0.37

	R^2	RMSE	F-ratio	R^2cv
PSA_S	0.393	0.25	44.702	0.35
PSA_S TSA	0.393	0.25	22.03	0.33
TSA ASA_S BSA_S	0.478	0.24	20.52	0.41
TSA PSA_S HalSA BenSA	0.397	0.26	10.89	0.31
TSA ASA_S BSA_S HalSA BenSA	0.490	0.24	12.47	0.41

From the results it can be seen that PSA_U descriptor gives a small improvement in R^2 and RMSE when split into ASA_U and BSA_U. If the descriptors HalSA and BenSA are added to the regression a negligible improvement is seen in R^2 and RMSE.

A two parameter model of PSA_S and TSA is inferior to the analogous equation made with PSA_U. Partition of PSA_S into ASA_S and BSA_S gives a vast improvement in R^2 with its value rising from 0.39 to 0.48. Addition of the HalSA and BenSA descriptor gives a small improvement to R^2 and RMSE. For these reasons we determine the best model of intestinal absorbance to be that which encompasses TSA, ASA_S, BSA_S, HalSA and BenSA.

While the scaled five parameter model offers only a 7% improvement in R^2 to that of a model of based only on PSA_U the inclusion of extra descriptors is merited by the lower number of false positive given.

5.4.3.3 Results Dataset III

This dataset represents the largest and most complex of all three intestinal absorbance datasets. Foscarnet was again removed from the regression, as surface area descriptors are not intended to model charged species.

The same regressions that were performed in 5.4.3.2 were repeated for this dataset. A one-parameter regression using PSA_U gives an R^2 value of 0.53 and an RMSE value of 0.19. A three-parameter model of TSA ASA_U and BSA_U gives only a small improvement in R^2 to 0.60 and a slight decrease in RMSE of 0.2 log units. A stepwise regression shows that after TSA, ASA_U and BSA_U are entered into the regression only BenSA and not HalSA was seen to be significant, Using these four descriptors the regression produced the following equation.

$$AbsT = 1.09 + 0.0002TSA - 0.0133ASA_U - 0.0021BSA_U + 0.0009BenSA \quad (5.10)$$

$N= 125, R^2 = 0.612, RMSE = 0.172, F\text{-ratio} = 47.351$

Similar models made using the scaled surface area descriptors exhibit similar properties to their unscaled counterparts. A simple regression of scaled PSA against AbsT gives an R^2 value of 0.50 and an RMSE of 0.19. Splitting of scaled PSA and inclusion of TSA improves the statistics. This equation is given below.

$$AbsT = 1.089 - 0.0005TSA - 0.0172ASA_S - 0.0016BSA_S \quad (5.11)$$

$N= 125\ R^2= 0.640\ RMSE = 0.165$ F- ratio 71.856

A five-parameter model was not produced as BenSA and HalSA are both seen to be insignificant at the 95% level. While the statistics of our scaled three parameter model are an improvement to a model produced using traditional PSA descriptor the results are still inferior to the equation produced using the Abraham LFER in which an R^2 value of 0.80 and RMSE value of 0.29 was obtained.[55] Although direct comparisons between the statistics of equations 5.10 and 5.11 and those of the LFER of Abraham are not possible as the Abraham LFER for this dataset predicted a first-order rate constant that was calculated from % absorption and not arcsine values as used within our study. For this reason a regression of the scaled surface area descriptors used in equation 5.11 were performed against the rate constant used by Abraham. The regression gave the following values.

$$Log\{ln[100/100-\%Abs)]\} = 0.572 + 0.0008TSA - 0.038ASA_S - 0.004BSA_S \quad (5.12)$$

$N = 125, R^2 = 0.67, RMSE = 0.343, F\text{-ratio} = 82.19$

The results of equation 5.12 are marginally better than those of 5.11 suggesting that the method of transforming the %absorbance data into a first-order rate constant stated by Abraham may be more suitable for modelling intestinal absorbance using surface area descriptors than our transformation into arcsin value methods, It should be noted that the dataset of Abraham is specially selected so as not to contain any %Absorption values of 0 or 100 %. The statistics of equation 5.12 are inferior to the LFER of Abraham published for this dataset, although the LFER of Abraham was reliant on experimentally observed descriptors.

5.4.3.4 Passive permeability Introduction

The models created for datasets I, II and III are notably less accurate than models created for simple non-biological systems such as water/octanol and water/chloroform partition as detailed in 4.2. The large difference in accuracy can be attributed in part to the uncertainty associated with biological data; this is often caused by the difficult experimental methods that are necessary to measure such properties. A further cause of error in our models is the presence of molecules in the dataset that are capable of active transport. While many studies have been reported and many molecules capable of active transport have been identified, such studies are not comprehensive enough to guarantee that none of the % absorbance values in our datasets are influenced by active transport.

Via the use of data obtained from *in vitro* methods both problems can be either significantly reduced or eliminated. While many methods for *in vitro* determination of absorption have been proposed, the parallel artificial membrane permeability assay (PAMPA) developed by Kansy *et al* [47] is the most suitable for our studies.

In vitro assays that imitate the multi-mechanism system of the intestine give results which are difficult to interpret in terms of individual mechanisms, where as PAMPA provides straightforward data on the prevalent mechanism for intestinal absorption, namely passive permeation. Many *in vitro* methods such as measures of permeability through Caco-2 monolayers are reliant on the labour intensive production of a monolayer, which reduces their efficiency and reproducibility. PAMPA assays are faster and highly reproducible, as

they do not require living cells as permeability is measured through an immobilised lipid membrane.

Huque *et al*[61] produced a model using the LFER of Abraham for the passive permeability of 40 small organic and drug like compounds measured using the PAMPA technique. The data is recorded as $\log P_o$ values. Huque obtained descriptors from experimentally observed values where possible, or else calculated them using the group contribution method of Platts. Partial least squares analysis was then used to produce the following equation. MLRA was not used as the descriptors B and V correlated heavily (R^2= 0.78).

$$\log P_o = -4.264 + 1.149E - 1.602S - 1.683A - 2.887B + 3.026V \qquad (5.13)$$

$N = 40, R^2 = 0.824$, RMSE = 0.836, $R^2_{CV} = 0.737$

5.4.3.5 Passive Permeability Results

Surface area descriptors were calculated for the 40 molecules used by Huque *et al*[61]. As none of our descriptors correlated highly, MLR was used to calculate the coefficient values. Two molecules were seen to be significant outliers and removed from the dataset, these were chlorpromazine and penbutolol: their removal was justified by the following reasons. Chlorpromazine has been noted as being an outlier and justifiably removed from previous surface area models (see 4.2). Penbutolol is removed, as the reliability of the base descriptors value is questionable due to the possibility of several intramolecular hydrogen bonds.

Simple models of the remaining 38 molecules using PSA_U gives an R^2 value of 0.55 and an RMSE of 1.21. If TSA is added to create a two-parameter equation the R^2 value rises to 0.794.

The Decoupling of PSA_U has negligible effect upon the model. The t-ratios for this three-parameter model show that of the two decoupled descriptors it is BSA_U that is the more dominant of the two. There is also seen to be a correlation between PSA_U and BSA_U of 0.99, explaining the similarity between the three and two parameter model. Addition of

the HalSA and BenSA descriptors to either the three or two parameter unscaled model has no effect.

In a one parameter regression with PSA_S a very poor model is produced with virtually no correlation between observed and calculated values (R^2>0.1), although in a two parameter model of PSA_S and TSA a vast improvement is seen in R^2 with its value rising to 0.744. The t-ratios show that the addition of TSA not only raises the accuracy of the model but also increases the significance of PSA_S. The de-coupling of PSA_S further improves the model with its R^2 value rising by 5%. From the t-ratios of the three-parameter model we see again that the hydrogen bond basic descriptor is the more dominant of the two, though unlike the unscaled counterpart, ASA_S is significant. The t-ratio values show ASA_S to contribute approximately half that of BSA_S, with both ASA and BSA reporting negative values. Equation 5.13 also shows the hydrogen bond basicity term B to be the more significant with the acidity term A contributing about 50% less. Further more, the most significant terms in each equation (5.13 and 5.14) are the size terms (V and TSA). The addition of the HalsA and BenSA descriptors to the model shows that the BenSA descriptor is insignificant at the 95% level, while HalSA has a small significance. The equation for the final four-parameter model is given below.

$$\log P_o = -4.5 + 0.015TSA - 0.133ASA - 0.046BSA - 0.008HalSA \qquad (5.14)$$

N= 38 R^2= 0.816 RMSE = 0.784 F-ratio = 36.61 R^2_{cv}= 0.763

Of all the equations produced 5.14 is seen to be the best although it should be stated that the R^2 of equation 5.14 is only 2% higher and RMSE 0.05 log units lower than a two parameter model of TSA and PSA_U.

5.4.4 Discussion

The t-ratios for each of the best surface area equations (based on R^2 values) for each dataset are given in table 5.10. The t-ratios for dataset I are not included because as stated before the set is so limited in size and chemical structure that any conclusions drawn may be unreliable.

Table 5.10: T-ratios of intestinal absorption and PAMPA permeability data

	T-Ratio				
	TSA	ASA_S	BSA_S	HalSA	BenSA
Dataset II	1.09	-4.02	-5.11	0.43	-1.15
Dataset III	3.21	-7.41	-6.18	-	-
Passive Permeability	7.57	-2.79	-8.366	-1.25	-

- Descriptor seen not to be significant within regression

Very little consistency can be seen in the significance of the descriptors between the intestinal absorbance models of datasets II and III. Both datasets show that the size term TSA will act to increase the Absorption of a molecule by the intestine, although the t-ratio of TSA for dataset II is bellow the significant value of 2 although more confidence is placed in the t-ratio value of dataset III as it is based on a substantially larger dataset. While this result may appear at first to be in opposition to Lipinski's[50] rule of five statement that any molecule displaying a Mw > 500 would exhibit poor absorption, it must be remembered that these models are generated from datasets of predominantly low molecular weight molecules, and while such a cut off could still occur the effects will not be manifested within the regression. ASA and BSA are both seen to be negative throughout, acting to reduced intestinal absorbance with the significance of ASA and BSA being almost equal in both models. This similarity in acidity and basicity was also reported in the Abraham LFER model of dataset III,[55] in which the hydrogen bond acidity term A has a coefficient value of -0.40 and B has a coefficient value of -0.51. This similarity in coefficient values and functions explains why the splitting of PSA has only a small effect on the models. Osterburg *et al*[62] also showed that N, O, N-H and O-H would have a negative effect on permeability. It should be mentioned that some studies such as that of Oprea[63] suggested that hydrogen bond donors should be more important that hydrogen bond bases as lipids in the cell membrane contain ester head groups that are capable of forming hydrogen bonds to donors but not hydrogen bond acceptors.

Further insight into the values of these descriptors can be gained if the mechanism for intestinal absorbance is considered. For molecules being absorbed by passive diffusion the route of permeation is believed to be via the transcellular pathway (across epithelial cells).[64] This transport can be considered a two-step process; a molecule must first cast off the water molecules that form its hydration sphere in order enter the lipid bilayer of the

cell membrane. The ease with which a molecule can shed this hydration sphere will greatly effect its absorbance and hence the reason that ASA_S and BSA_S both report negative values. The second step of absorption is the molecules transport through the inner cell before crossing the cell membrane again on its exit. In order for this second transport to take place it would be expected that a molecules would need to exhibit an affinity for lipids. The t-ratios exhibited for datasets II and III reflect known lipophilic properties such as a positive size term and negative hydrogen bond basicity (as reported for the model of $logP_{oct}$ reported in 4.2.2).

The descriptors HalSA and BenSA are seen to be insignificant either via stepwise regression or t-ratio values of less than 2, for this reason the best model for intestinal absorption and passive permeability is a three parameter model using the descriptors TSA, ASA_S and BSA_S.

The t-ratios of the passive permeability model are similar to those of the Abraham LFER of this dataset published by Huque *et al.*[61] Negative values are reported for the hydrogen bond acidity and basicity terms acting again to hinder permeability across the lipid bilayer. The hydrogen bond basicity term is seen to be of greater importance than the hydrogen bond acidity term, a result also displayed by the model of Huque and in agreement with the conclusion of Oprea.[63]

5.5 Cell Permeation

5.5.1 Introduction

Many important biological/pharmaceutical properties are reliant on a molecule's ability to penetrate across cell wall membranes. While previous models such as gastro-intestinal absorbance include for cell permeation, they are affected by other factors. A model that predicts only cell permeation would prove very informative.

The rate of permeation of compounds into the giant algal cells *Chara certaophylla* and *Nitella*, have been reported be Collander *et al.*[65,66] From these rates of permeation Collander reported correlations with water/ether and water/olive oil partition coefficients. A subset of the data for *Chara certaophylla* cells was further studied by Raevsky and Schaper[67] who found that a molecules hydrogen bond capacity was influential to its permeability.

Platts *et al*[68] applied the LFER of Abraham to Collander's permeation data, and created separate models for *Chara certaophylla* and *Nitella* cells. These models were created using descriptors generated via the group contribution of Platts. For a dataset of 37 permeation values into *Chara certaophylla* cells, a model with an R^2 of 0.962 was created this model was seen to be dominated by the hydrogen bond acidity descriptor A. A similar model of based on the rates of permeation for 63 *Nitella* cells gave an R^2 value of 0.881. The two datasets were combined to create a generic model of cell permeability this model was also seen to be highly accurate with an Sd of 0.437.

We have chosen to take the two datasets of Collander and create similar models to those produced by Platts *et al* using our surface area methods. In addition to the permeability data for compounds into living Nitella cells colander also reported uptake data in to dead cells, this data was also modelled using our methods.

5.6.2 Methods

Surface area descriptors were generated for the datasets of the permeability of 37 molecules into *Chara certaophylla* cells and 63 molecules into *Nitella* cells, using the methods outlined in 4.9. For each dataset three models were produced, these models used the following descriptors

1. TSA, PSA_U
2. TSA, ASA_U, BSA_U, HalSA and BenSA
3. TSA, ASA_S, BSA_S, HalSA and BenSA

5.5.3.1 Results Chara Ceratophylla Cells $logK_{cc}$

Initial analysis showed thiourea to be an outlier, as it is well predicted within the models of Platts, (Obs = -2.11 Calc = 2.32) it can be assumed that the observed value is not the source of error. The error is most likely caused by the incorrect assignment of scaling factors by the algorithm as no values occur for thiourea. Instead the polar atoms are scaled with values taken from thioamides. The differences between the A and B values of thiourea and thioamides can be seen if we compare experimentally observed A and B values of thiourea and thioacetamide.[69] Thiourea has A and B value of 0.77 and 0.87 respectively while thioacetamide has values of 0.58 and 0.64. For these reasons we omit thiourea from our models. In the study of Platts *et al* lactamide was a noted outlier and omitted without reason. Using our methods lactamide is well predicted, so it is kept in the regression. The statistical analysis of our models is given in table 5.11 along with the results from the LFER model produced by Platts.

Table 5.11: Results of surface area models of permeation into Chara ceratophylla Cells

Descriptors	N	R^2	RMSE	R^2_{CV}
TSA PSA	36	0.44	0.90	0.30
TSA ASA_U BSA_U HalSA BenSA	36	0.58	0.82	0.43
TSA ASA_S BSA_S HalSA BenSA	36	0.86	0.47	0.82
LFER	36	0.96	0.25	0.94

It is evident from table 5.11 that the scaled five parameter model is the best surface area model, although the LFER method is the most accurate of all.

Analysis of the t-ratios for the five parameter model show that the most dominant descriptor is ASA_S which acts to lower the rate of permeation of a molecule. The next most importance term is BSA_S that also acts to reduce cell uptake. Exactly the same importance and signs of descriptors were seen for the analogous hydrogen bond acidity and basicity descriptor terms in the study of Platts. The size descriptor TSA is seen to be the third most significant descriptor acing to increase cell permeability.

5.5.3.2 Results Nitella Cells $logK_{nit}$

Again Thiourea was very poorly predicted especially in models where scaled ASA_S and BSA_S descriptors were applied again we choose to omit it from the regression. For this datasets the exact same pattern was seen in the accuracy of the different models, with a simple two-parameter model being the least accurate and the scaled five-parameter model being the superior. As with the model of $logK_{cc}$ a difference of approx 10% in R^2 was seen between the five parameter scaled model and the LFER.

Table 5.12: Results of surface area models of permeation into Nitella Cells

Descriptors	n	R^2	RMSE	R^2cv
TSA PSA	63	0.32	1.12	0.23
TSA ASA_U BSA_U HalSA BenSA	63	0.58	0.90	0.50
TSA ASA_S BSA_S HalSA BenSA	63	0.81	0.60	0.77
LFER	63	0.88	0.46	0.83

The t-ratios for the five parameter scaled model shows them to be remarkably similar to those of the model of *Chara certaophylla* cells. The ASA_S descriptor is the most influential followed by BSA_S with both acting to reduce cell uptake, Again TSA gives a smaller contribution and acts to increase cell uptake.

5.5.3.3 Results Dead Nitella Cells $logK_{nit}$

Both Collander and Platts suggested that uptake into dead cells is governed entirely by molecular weight and properties such as charge and hydrogen bond acidity are not important. In order to evaluate this in addition to the normal models a one parameter model with TSA was also produced. The results are given in the table below.

Table 5.13: Results of surface area models of permeation into dead Nitella Cells

Descriptors	n	R^2	RMSE	R^2cv
TSA PSA	64	0.91	0.04	0.89
TSA ASA_U BSA_U HalSA BenSA	64	0.92	0.04	0.91
TSA ASA_S BSA_S HalSA BenSA	64	0.92	0.04	0.90
TSA	64	0.89	0.04	0.88
LFER V only	64	0.93	0.04	0.93

The results in table 5.13 show that the uptake by dead cells is indeed governed almost entirely by molecular size with the effects of hydrogen bond acidity and basicity being virtually inconsequential, with a difference of only 0.03 between the five-parameter models and that of the model of just TSA.

It is interesting to note that for $logK_{Nit\ Dead}$ where the descriptors ASA and BSA have very little significance thiourea is not reported an outlier confirming our hypothesis that the error for this molecule is due to miscalculation of ASA_S and BSA_S.

5.5.3.4 Results Combined Chara ceratophylla and Nitella Cells $logK_{gen}$

The broad similarities between the coefficient values for the models of Chara ceratophylla and Nitella, and the similarity in rates of permeation for the 27 molecules that occur in both datasets, indicates that the systems are similar enough that the two separate datasets can be combined to create a model of generic cell permeation. Thiourea was again removed from the regression. The combined dataset was remodelled using the descriptors TSA ASA_S BSA_S HalSA and BenSA. This regression yielded the following model.

$LogK_{gen}$ = -3.013 +0.002TSA –0.218ASA_S -0.018BSA_S +0.012HalSA –0.019 (5.14)

N= 98, R^2 = 0.821, RMSE=0.551, F ratio = 84.23, R^2_{cv} = 0.79

While the accuracy of this model is less than those of the separate models the reduction is accuracy is probably the result of the wider range of compounds. The statistical analysis of a number of other models made using the combined dataset is given in table 5.14.

Table 5.14: Results of surface area models of permeation into Chara ceratophylla and Nitella Cells

Descriptors	n	R^2	RMSE	R^2_{cv}
TSA PSA	98	0.39	1.00	0.34
TSA ASA_U BSA_U HalSA BenSA	98	0.59	0.83	0.55
TSA ASA_S BSA_S HalSA BenSA	98	0.82	0.55	0.79
LFER V only	100	0.89	0.44	0.87

5.6.4 Discussion

The t-ratios for each of the models created using the five descriptors with hydrogen bond scaling applied are given in table 5.15.

Table 5.15:t-ratio values of cell uptake models

	Chara ceratophylla	Nitella	Dead Nitella	Combined model
Term	t-ratio	t-ratio	t-ratio	t-ratio
TSA_S	2.78	1.47	-21.59	2.56
ASA_S	-11.09	-12.23	-1.79	-15.96
BSA_S	-8.40	-9.64	-3.34	-12.47
HalSA	1.47	0.51	-1.72	1.16
BenSA	-2.51	-1.64	-1.65	-2.69

The t-ratios show that for all models of uptake by living cells the ASA_S descriptor is the most influential. The second most influential descriptor is seen to be BSA_S. From these findings it can be concluded that the cell interior is much less hydrogen bond basic than

bulk water and to a lesser extent the cell interior is less acidic than bulk water. The positive TSA value can be attributed to cavity effects caused by the highly dense nature of bulk water, although these effects are far less dominant than those of hydrogen bond acidity and basicity for living cells.

The descriptors HalSA and BenSA are seen to be of little importance with HalSA acting to increase cell uptake and BenSA acting to decrease cell up take. Due to the low importance of both these descriptors it is difficult to conclude much about their role in the mechanism of cell uptake the low t-ratio values of HalSA throughout imply that it should be removed from predictive models of cell uptake.

5.6 Conclusions

5.6.1 Conclusions -Uptake of volatile organic compounds by plants

Models have been produced for $\log K_{MXa}$ and $\log K_{MXw}$ for Lycopersicon (tomato fruit), while the statistics offered by these models were lower than those for the models proposed by Platts *et al* created using the LFER of Abraham, the models were still very good. Via analysis of the t-ratios obtained from these models conclusions were drawn about the factors governing the mechanism of partition. These conclusions were in agreement of those from previously published work and were physically sound.

5.6.2 Conclusions - Partition into Biological liquids and tissues of vapours and biological liquids

Models have been produced for a wide variety of biological liquids and tissues. These models, while not as accurate as those produced using the LFER of Abraham, are still good with R^2 values as high as 0.903 for the solubility of 42 molecules in human muscle tissue. The models offer a distinct advantage over those published by Abraham as they are based entirely on theoretically calculated descriptors and not reliant experimentally observed descriptors.

From the t-ratios of the descriptors used within this study it was seen that generally biological tissues displayed more lipophilic than hydrophilic properties. It was also seen that biological fluids such as blood and more so plasma shared many properties with water.

Analysis of the descriptors over all models revealed that for biological fluids and tissues the most significant and influential of the descriptors are ASA_S and TSA. Of a lesser importance, but still requiring consideration, are the BenSA and ClSA descriptors. The BSA descriptor is almost completely redundant and is seen to be the least significant within all models of biological tissues or fluids.

5.6.3 Conclusions - Blood Brain Barrier

Three separate models of blood brain barrier have been produced using scaled and un-scaled polar surface area descriptors. The best models produced use the descriptors TSA ASA_U BSA_U and BenSA. Although the models are seen to be only a small improvement over models in which just PSA_U is used as a descriptor.

Attempts to apply our scaling factors gave no improvement to the models but instead reduced the quality of the statistics. This increased error is possibly due to specific functional groups having unique properties within blood brain partition, properties that are not accounted for by traditional definitions such as hydrogen bond acidity and basicity. The traditional definition of PSA is seen to be superior, as it dose not attempt to discriminate between different functional groups.

5.6.4 Conclusions - Intestinal Absorption

It was seen that our method of generating PSA from a single low energy conformation produced models of intestinal absorption of equal quality to those which employed the dynamic PSA descriptor calculated by Palm. If a five parameter unscaled model is created the resultant model is of much higher accuracy than that of a model of PSA_d.

The intestinal absorbance models created from datasets II and III, show that scaling and partition of PSA_U gives a small improvement to R^2 and RMSE. This improvement is most noticeable for dataset III, the larger improvement is most likely due to the wider range of function groups present in dataset III. Models produced for passive permeability data were seen to be as accurate as those made using the LFER of Abraham. Analysis of the t-ratios of this model shows that the coefficients calculated for these models are physically realistic and similar to those generated from other studies.

Previous studies of intestinal absorbance have concluded that PSA_U is a useful tool for predicting intestinal absorbance as it measures the hydrogen bonding capacity of molecules. Our results suggest that PSA_U is more specifically a measure of a molecules hydrogen bond basicity as BSA_U and PSA_U in datasets II and III, are seen to correlate > 0.99. Also the partitioning of PSA_U to ASA_U and BSA_U yielded only small improvement, suggesting that where traditional definitions of PSA are used the surface area of Nitrogen and Oxygen atoms is of greater importance. Where scaling factors are applied the significance of the PSA_S descriptor drops off. Scaled descriptors only produce models comparable to their unscaled analogues when PSA_S is partitioned into ASA_S and BSA_S. Individually neither PSA_S, ASA_S or BSA_S can accurately produce models of intestinal absorbance.

5.6.5 Conclusion Cellular Uptake

The models produced confirm that our surface area descriptors are capable of producing models of cell permeation. Via comparison of surface area models created using different combinations of descriptors it is evident that splitting and scaling of PSA is essential to correctly model this processes. The equations produced for the models of Chara ceratophylla and Nitella Cells show great similarities in the dependence of the surface area descriptors with ASA_S being the most dominant descriptor. The similarities between these models are so strong that it is possible to combine both datasets and produce a generic model of cell permeation. The model produced for dead Nitella cells showed that only the TSA descriptor was influential and that the process of molecular uptake for dead cells is governed almost entirely by molecular size. These findings are in conclusion with those of previous studies.

5.7 Future work

The HalSA descriptor was seen to be less effective than ClSA in specific models of absorption of gases into biological tissues produced in 5.2.3, This preference for ClSA was seen in datasets where polarisabilty was seen to be important and the datasets contained a large number of different halogen atoms. The result implied that the accuracy of HalSA may be improved if specific scaling factors were added to HalSA to account for the varying polar/polarisable of different halogen atoms.

The encouraging results given by cellular uptake model and the models for inorganic partitions of $\log P_{oct}$ and $\log P_{CHCl}$ produced in 4.10 suggests that the scaled PSA method is capable of producing an accurate model for the cellular uptake of inorganic molecules

5.8 References

1. D.L. Dowdy, T.E. McKone, *Environ. Toxicol. Chem.*, **1997**. 16. 2448.
2. A. Sabljic, H. Gusten, J. Schonherr, M. Riederer, *Environ. Sci. Technol.*, **1990**. 24. 1321.
3. S. Paterson, D. Mackay, E. Bacci, D. Calamari, *Environ. Sci. Technol.*, **1991**. 25. 866.
4. S.L. Simonich, R.A. Hites, *Environ. Sci. Technol.*, **1995**. 29. 2905.
5. S. Merk, M. Riederer, *J. Exp. Bot.*, **1997**. 48. 1095.
6. R. Keymeulen, G. DeBruyn, H. VanLangenhove, *J. Chromatogr. A*, **1997**. 774. 213.
7. R.H.A. Brown, J.N. Cape, J.G. Farmer, *Chemosphere*, **1998**. 36. 1799.
8. B. Welke, K. Ettlinger, M. Riederer, *Environ. Sci. Technol.*, **1998**. 32. 1099.
9. J.A. Platts, M.H. Abraham, *Environmental Science & Technology*, **2000**. 34. 318.
10. M.H. Abraham, J. Andonianhaftvan, G.S. Whiting, A. Leo, R.S. Taft, *J. Chem. Soc.-Perkin Trans. 2*, **1994**. 1777.
11. J.A. Platts, D. Butina, M.H. Abraham, A. Hersey, *J. Chem. Inf. Comput. Sci.*, **1999**. 39. 835.
12. J. Schonherr, M. Riederer, *Rev. Environ. Contam. Toxicol.*, **1989**. 108. 1.
13. J.J. Slaski, D.J. Archambault, X. Li, *Report for Air Research UserGroup, Alberta Environment*, **2000**. ISBN 0-7785-1228-2.
14. V. Fiserovabergarova, *Model. Inhilation Exposure Vap.*, **1983**. 1. 3.
15. V. Fiserovabergerova, M.L. Diaz, *International Archives of Occupational and Environmental Health*, **1986**. 58. 75.
16. P.K. Weathersby, L.D. Homer, *Undersea Biomed. Res.*, **1980**. 7. 277.
17. A. Feingold, *Anesth. Anal. (N.Y)*, **1976**. 55. 593.
18. A. Sato, T. Nakajima, *Arch. Environ. Health*, **1979**. 34. 69.
19. M.H. Abraham, M.J. Kamlet, R.W. Taft, R.M. Doherty, P.K. Weathersby, *Journal of Medicinal Chemistry*, **1985**. 28. 865.
20. M.H. Abraham, P.K. Weathersby, *J. Pharm. Sci.*, **1994**. 83. 1450.
21. G. Pezzango, S. Ghittori, M. Imbriani, *Giorn. Ital. Med. Lav.*, **1985**. 7. 17.
22. L. Perbellini, F. Brugnone, D. Caretta, G. Maranelli, *British Journal of Industrial Medicine*, **1985**. 42. 162.
23. M.J. Kamlet, D.J. Abraham, R.M. Doherty, R.W. Taft, M.H. Abraham, *Journal of Pharmaceutical Sciences*, **1986**. 75. 350.
24. M.H. Abraham, *J. Am. Chem. Soc*, **1979**. 75. 350.
25. M.H. Abraham, J.R.M. Gola, *Unpublished Work*, **1997**.
26. M.H. Abraham, G.S. Whiting, R.M. Doherty, W.J. Shuely, *J. Chromatogr.*, **1991**. 58. 213.
27. M.H. Abraham, *J. Chromatogr.*, **1993**. 64. 95.
28. A. Reichel, D.J. Begley, *Pharmaceutical Research*, **1998**. 15. 1270.
29. D.E. Clark, *Comb. Chem. High Throughput Screen.*, **2001**. 4. 447.
30. G.W. Goldstein, A.L. Betz, *Sci. Am.*, **1986**. 255. 74.
31. O.G. Mouritsen, K. Jorgensen, *Pharm. Res.*, **1998**. 15. 1507.
32. W.M. Pardridige, *J. Neurochem*, **1998**. 70. 1781.
33. D.J. Begley, *J. Pharm. Sci.*, **1996**. 48. 136.
34. H. Vandewaterbeemd, M. Kansy, *Chimia*, **1992**. 46. 299.
35. J. Kelder, P.D.J. Grootenhuis, D.M. Bayada, L.P.C. Delbressine, J.P. Ploemen, *Pharm. Res.*, **1999**. 16. 1514.

36. D.E. Clark, *Journal of Pharmaceutical Sciences*, **1999**. 88. 815.
37. J.A. Platts, M.H. Abraham, Y.H. Zhoa, A. Hersey, L. Ijaz, D. Butina, *Eur. J. Med. Chem*, **2001**. 36. 719.
38. F. Lombardo, J.F. Blake, W.J. Curatolo, *J. Med. Chem.*, **1996**. 39. 4750.
39. M.H. Abraham, H.S. Chadha, R.C. Mitchell, *Drug. Des. Discov.*, **1995**. 13. 123.
40. U. Norinder, P. Sjoberg, T. Osterberg, *Journal of Pharmaceutical Sciences*, **1998**. 87. 952.
41. F. Lombardo, J.F. Blake, W.J. Curatolo, *Journal of Medicinal Chemistry*, **1996**. 39. 4750.
42. M.H. Abraham, H.S. Chadha, R.C. Mitchell, *Journal of Pharmaceutical Sciences*, **1994**. 83. 1257.
43. T. Salminen, A. Pulli, J.J. Taskinen, *Pharm. Biomed. Anal.*, **1997**. 15. 469.
44. M. Feher, E. Sourial, J.M. Schmidt, *Int. J. Pharm*, **2000**. 201. 239.
45. P. Artursson, K. Palm, K. Luthman, *Advanced Drug Delivery Reviews*, **1996**. 22. 67.
46. P. Artursson, R.T. Borchardt, *Pharm. Res.*, **1997**. 14. 1655.
47. M. Kansy, F. Senner, K. Gubernator, *Journal of Medicinal Chemistry*, **1998**. 41. 1007.
48. S.D. Kramer, *Pharmaceutical Science and Technology Today*, **1999**. 2. 373.
49. D.E. Clark, S.D. Pickett, *Drug. Discov. Today.*, **2000**. 5. 49.
50. C.A. Lipinski, F. Lombardo, B.W. Dominy, P.J. Feeney, *Advanced Drug Delivery Reviews*, **1997**. 23. 3.
51. K. Palm, P. Stenberg, K. Luthman, P. Artursson, *Pharmaceutical Research*, **1997**. 14. 568.
52. D.E. Clark, *Journal of Pharmaceutical Sciences*, **1999**. 88. 807.
53. O.A. Raevsky, V.I. Fetisov, E.P. Trepalina, J.W. Mcfarland, K. Schaper, *Quant. Struct. Act. Relat.*, **2000**. 19. 2000.
54. M.D. Wessel, P.C. Jurs, J.W. Tolan, S.M. Muskal, *Journal of Chemical Information and Computer Sciences*, **1998**. 38. 726.
55. M.H. Abraham, Y.H. Zhao, J. Le, A. Hersey, C.N. Luscombe, D.P. Reynolds, G. Beck, B. Sherborne, I. Cooper, *Eur. J. Med. Chem.*, **2002**. 37. 595.
56. V.S. Chungi, D.W.A. Bourne, L.W. Ditter, *J. Pharm. Sci.*, **1979**. 68. 1552.
57. P.K. Dudeja, S.A. Torania, H.M. Said, *American Journal of Physiology-Gastrointestinal and Liver Physiology*, **1997**. 35. 272.
58. J.F. Westphal, A. Deslandes, J.M. Brogard, C. Carbon, *Journal of Antimicrobial Chemotherapy*, **1991**. 27. 647.
59. B.G. Reigner, W. Couet, J.P. Guedes, J.B. Fourtillan, T.N. Tozer, *Journal of Pharmacokinetics and Biopharmaceutics*, **1990**. 18. 17.
60. N. RuizBalaguer, A. Nacher, V.G. Casabo, M. Merino, *Antimicrobial Agents and Chemotherapy*, **1997**. 41. 445.
61. F.T.T. Huque, K. Box, J.A. Platts, *AWAITING PUBLICATION*, **2004**.
62. U. Norider, T. Osterberg, *J. Pharm. Sci.*, **2001**. 90. 1076.
63. T.I. Opera, *J. Comput. Aided Mol. Des.*, **2000**. 14. 251.
64. S.D. Kramer, *Pharm. Sci. Technol. Today.*, **1999**. 2. 373.
65. R. Collander, H. Barlund, *Acta. Bot. Fenn.*, **1933**. 11. 1.
66. R. Collander, *Physiologia Plantarum*, **1954**. 13. 363.
67. O.A. Raevsky, K.J. Schaper, *Eur. J. Med. Chem.*, **1998**. 33. 799.
68. J.A. Platts, M.H. Abraham, A. Hersey, D. Butina, *Pharm. Res.*, **2000**. 17. 1013.
69. M.H. Abraham, P.L. Grellier, D.V. Prior, P.P. Duce, J.J. Morris, P.J. Taylor, *J. Chem. Soc. Perkin Trans. 2*, **1989**. 6. 699.

Chapter 6. Industrial properties and green solvents

6.1 Fluorophilicity

6.1.1 Introduction

The rapidly increasing worldwide demand for environmentally friendly chemical properties and processes has resulted in an explosion of interest in environmentally friendly 'green' chemistry particularly catalysis. One popular area of green chemistry is fluorous biphasic catalysis;[1-4] this method employs environmentally benign solvents, produces a high product yield and provides an efficient catalyst recycling system. Horvath and Rabai first reported catalysis performed in a fluorous biphase system in 1994.[5] A fluorous biphasic system consists of a fluorous phase such as perfluoro(methylcyclohexane) containing a preferentially fluorous phase soluble catalyst, and a second phase, which may be any organic or inorganic solvent with limited solubility in the fluorous phase. The two phases have limited miscibility at room temperature but homogenise at higher temperatures, allowing reactants in the organic phase to interact with the catalyst in the fluorous phase at high temperature, with the advantage that upon cooling the homogenised system will separate in to its component phases, efficiently partitioning the reaction products into the organic phase and the catalyst into the fluorous phase.

The most important factor in designing such a system is the solubility of products, reactants and catalyst in the fluorous phase. The ability to predict this directly from chemical structure would be of great importance in the design of new fluorous biphasic systems, allowing unsuitable molecules to be identified and eliminated at an early stage in the design process, saving both time and money.

The tendency of a molecule to dissolve in fluorous media may be quantified by its fluorophilicity, which is determined by taking the natural logarithm of a molecule's partition coefficient P, between fluorous and organic layers.[6] Within this study the standard system proposed by Rocaboy *et al*[7] shown in equation 6.1 has been employed

$$\ln P = \ln\left[\frac{c(CF_3C_6F_{11})}{c(CH_3C_6H_5)}\right] \quad T = 298K \qquad (6.1)$$

Kiss *et al*[8] produced a model of fluorophilicity for 59 fluorinated organic molecules using a neural network combination of eight descriptors chosen from a pool of nearly 100 descriptors. The initial pool of descriptors included properties such as electrostatic potential, HOMO and LUMO energies, and weighted holistic invariant molecular (WHIM) descriptors. The surface area of the molecule and the distribution of fluorine were found to be significant. The study showed that fluorophilicity was increased when the fluorine atoms were on the exterior of the molecules and capable of interacting with the fluorous phase. Kiss *et al*'s final model stated that a molecules fluorous content had little significance upon its fluorophilicity, in conflict with the established view of the factors that determining fluorophilicity. This was assigned to the lack of molecules with little fluorine content within their dataset.

Huque *et al*[9] produced a model for the prediction of fluorophilicty for 91 organic molecules, this dataset *al*so contained more molecules with no fluorine atoms than the dataset of Kiss *et al*. This model employed a modified version of linear free energy relationship (LFER) of Abraham and coworkers.[10] Huque *et al* added an additional sixth descriptor F, the fluorine content of the solute. The descriptors for this study were not experimentally observed but calculated using the group contribution of Platts.[11] Stepwise regression demonstrated that B was insignificant at the 95% level, and was removed from the model. The coefficients of this model revealed that the most influential descriptor upon fluorophilicity is F indicating that molecules with high fluorine content would preferentially be drawn into the fluorous phase.

Duchowicz *et al*[12] proposed a model of fluorophilicity based on multivariate regression of very simple topological molecular descriptors, which were obtained from counting the number of atoms and bonds in the molecule. When applied to the same dataset of 91 organic molecules used in the study of Huque[9] Duchowicz produced a model of equal quality using 23 descriptors. Duchowicz stated that while the model of Huque was better

as it relied on only five descriptors the ease at which his proposed descriptors could be calculated should not be overlooked.

Recently de Wolf *et al*[13] applied the universal lipophilicity model based on the mobile order and disorder (MOD) solution theory to predict the partition coefficients for 88 molecules in either PFMCH/toluene or FC-72/benzene. Prediction required knowledge of molecular volume and modified non-specific cohesion parameter of the solute; de Wolf also detailed methods in which these properties could be easily calculated. The model showed that extending the perfluoroalkyl tails on a given substance would not automatically result in higher partition coefficients.

While these methods have been seen to work well for organic compounds, no predictive models occur for organometallic molecules containing transition metals. The ability to predict the fluorophilicity of transition metal complexes is of great importance when one considers the number of organometallic catalysts used in fluorous biphasic catalysis.

Following the successes of the scaled PSA approach used in Chapter 4.10 to predict partition properties of Platinum compounds models were constructed for calculation of fluorophilicity for organometalic molecules.

6.1.2 Method

The same set of 98 organic and fluorous molecules used in the study of Huque *et al*,[9] originally taken from Gladysz's online database,[14] was initially used. Eight transition metal complexes, not used in any previous study, were taken from the same source – these are reported in Table 6.1 For the 98 organic molecules structures were generated via CORINA and energy minimized using AM1 methods.

Table 6.1: Structures and fluorophilicities of transition metal complexes

Code	Name [a]	ln*P*
Tm1	$[\{R_{f6}(CH_2)_2\}_3P]_2Ir(Cl)(CO)$ [b]	5.81
Tm2	$[\{R_{f6}(CH_2)_2\}_3P]_2NiCl_2$	4.41
Tm3	$[R_{f10}(CH_2)_2C_5H_4]Mn(CO)_3$	0.59
Tm4	$[R_{f10}(CH_2)_2C_5H_4]Rh(CO)_2$	-0.22
Tm5	$[R_{f10}(CH_2)_2C_5H_4]Rh(CO)[P\{(CH_2)_2R_{f6}\}_3]$	3.38
Tm6	$[R_{f10}(CH_2)_2C_5H_4]_2Fe$	2.99
Tm7	$[R_{f6}(CH_2)_2C_5H_4]_2ZrCl_2$	3.03
Tm8	$[R_{f6}(CH_2)_2C_5H_4]_2Zr(CH_3)_2$	1.95

[a] $R_{fn} = (CF_2)_{n-1}CF_3$

[b] Structure obtained from X-ray crystallography

Structures for transition metals were obtained from the study of Platts *et al*, who employed Morokuma's ONIOM method,[15] as implemented in Gaussian03,[16] running on the UKCCF's Columbus facility. In this method, the transition metal core is treated using a quantum mechanical method such as DFT, while outer regions are treated using a molecular mechanical method. Following a series of tests and comparisons to structures obtained from X-ray crystallography, the B3LYP/Lanl2DZ method was found to be the best method for the QM region and AMBER for the MM. Force constants for bonds and angles not contained in the standard AMBER force field, *i.e.* those associated with the transition metal center, were set to zero such that they make no contribution to the ONIOM energy or geometry.

From these optimized structures, PSA descriptors were calculated following the methods set out in 4.9. The halogen surface area descriptor was modified to include only the surface area of fluorine atoms, denoted FSA. A further descriptor, MetalSA, was defined as the exposed surface area of the central metal ion.

This raises a further complication, since calculation of PSA requires van der Waals radii of all atoms, and no such radii are available from standard sources for Ni, Ir, Rh, Mn, Fe, or Zr. In order to address this problem an in-house C-program was used to search for the

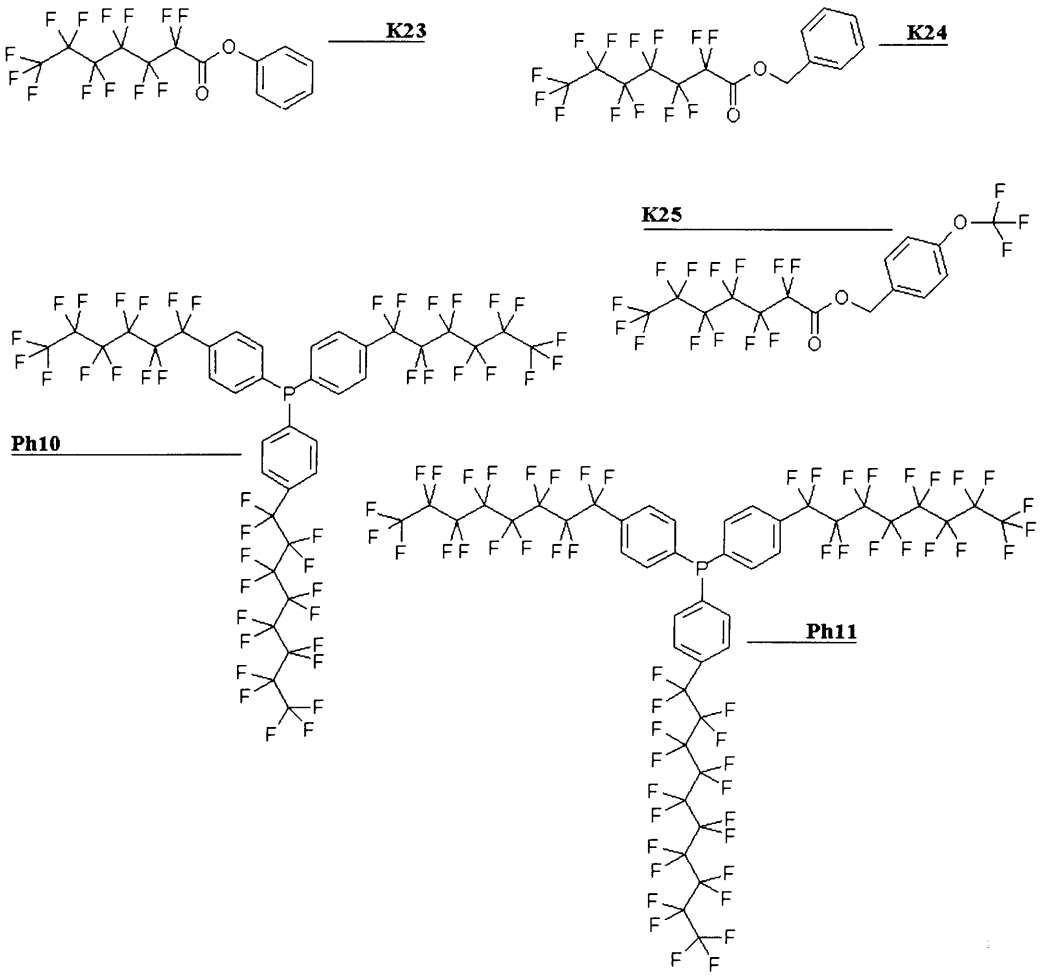
point on the 0.001 e.au^{-3} isodensity surface closest to the metal center, this distance was then taken to be the vdw radii and relevant modifications were made to MOLVOL.

6.1.3 Results and Discussion

Initial regression of the 98 ln *P* values for organic molecules against simple polar and total surface areas (TSA & PSA_U) gave poor results ($R^2 = 0.41$, RMSE = 2.01), indicating that such a model is inadequate for the task in hand. This agrees with our findings for more conventional partition coefficients such as $logP_{oct}$, where such simple models were shown to be inadequate. Introduction of FSA yields a marked improvement, doubling the R^2 to 0.84 and halving RMSE to 1.05. This improvement is much larger than was seen on adding the weight-fraction of F as a descriptor in the LFER model of Huque, supporting Kiss *et al*'s findings that the distribution of fluorine is as important as the total amount in determining fluorophilicity. Further partitioning and scaling of PSA descriptors results in only small statistical gains, such that the best five parameter model for these 98 data has $R^2 = 0.88$ and RMSE = 0.93.

As noted in previous efforts to model this dataset,[8,9] several molecules appear as outliers, whether for statistical or physical reasons. One class of molecules, denoted K23, K24, and K25 (shown in figure 6.1), do not show consistent physical properties: for instance K23 and K24 differ only by a single CH_2 group, but their ln*P* values differ by more than 1.5 units. These three compounds were therefore omitted from all further analyses. In this study, two further molecules Ph10 and Ph11 (shown in figure 6.1) are also statistical outliers. Both are tri-phenyl phosphines, for which very little data was available during the scaling process, set out in 3.1.3.1, consequently there is less confidence for these molecules than other classes of compounds. These two compounds are also therefore omitted. However, it is encouraging that one class of outlier in the LFER study is well modeled here, *i.e.* that with compounds containing more than one fluorous chain attached to a single aromatic ring. It was also possible to include three silicon-containing molecules that were omitted from the LFER study, as no fragmental Abraham values exist for silicon.

Figure 6.1: structures of outliers



With the removal of these outliers, the same models as described above were applied. Again, the simple TSA + PSA_U model is poor (R^2 = 0.47, RMSE = 1.92), while introduction of FSA gives an excellent three-parameter model, with R^2 = 0.93 and RMSE = 0.71. As before, splitting and scaling the PSA descriptor gives small improvements, such that our final five-parameter model for this organic dataset is the following:

$$\ln P = -0.723 - 0.001\ TSA - 0.0909\ ASA_S - 0.0146\ BSA_S + 0.0209\ FSA - 0.0115\ BenSA \qquad (6.2)$$

$N = 93,\ R^2 = 0.944,\ RMSE = 0.638,\ F = 291.15,\ R^2_{CV} = 0.935$

This model is of almost identical accuracy as the previously published LFER model, confirming the suitability of the surface area descriptors used. It is also in full agreement regarding the physical significance of descriptors, since by far the most significant term is FSA (t-ratio = +26.6), indicating that exposed fluorine atoms act to increase ln *P*. TSA has less than half this significance (t = -12.6), such that larger solutes prefer the organic to

fluorous phase. All other t-ratios are small and negative, revealing that molecules containing H-bonding and/or polar groups also prefer the organic phase, as might be expected when comparing the physical properties of toluene and perfluoro (methylcyclohexane).

Thus, equation 6.2 provides an alternative model of equal accuracy to previous models, and could be used as independent corroboration for prediction of fluorophilicity of organic compounds. However, the use of surface area properties opens up the possibility of modeling the fluorophilicity of metal complexes. Fluorophilicities of eight transition metal complexes were available from Herrera *et al*[17] and Gladysz's online database (Table 6.1). Three further transition metal complexes $[\{R_{f6}(CH_2)_2\}_3P]_3$**Rh**Cl, $[\{R_{f8}(CH_2)_2\}_3P]_3$**Rh**Cl and $[\{R_{f6}(CH_2)_2\}_3P]_3$**Rh**(H)(CO) were available from the online database but not used within our models as preliminary studies showed the size and flourous content of these molecules to be so much larger than those of the other molecules in the dataset, such that the values calculated for descriptors would fall outside the range of the model.

Seven of the eight transition metal complexes reported in Table 6.1 were optimized using the same ONIOM(B3LYP/Lanl2DZ:AMBER) method as detailed by Platts *et al.* The eighth compound, Tm1, could not be optimized due to a limitation in the Gaussian03 package which prevents use of AMBER for third-row transition metals such as Ir. Fortunately, a low temperature crystal structure of this complex has been reported [18], although some positional disorder is present in the fluorous chains. This structure was used without modification for all subsequent predictions. Exposed surface areas of these structures were then calculated, initially using a standard radius of 2Å for all transition metals.

Initial models incorporated the transition metal complexes into the organic dataset used to develop equation 6.2. As expected, a simple TSA + PSA_U model is poor (R^2 = 0.48), but introduction of fluorine surface area yield a reasonably accurate three parameter model (R^2 = 0.88, RMSE = 0.95). Still greater accuracy can be obtained by adding in MetalSA, while splitting and scaling of the PSA descriptor has a small beneficial effect on the model. T-ratios revealed that BSA_S is insignificant within the regression, in agreement with Huque *et al.* Since neither toluene nor perfluoro(methylcyclohexane) displays any

hydrogen bond acidity, hydrogen bond basicity has no influence into which phase a molecule will partition, hence BSA_S is omitted from the model yielding equation 6.3.

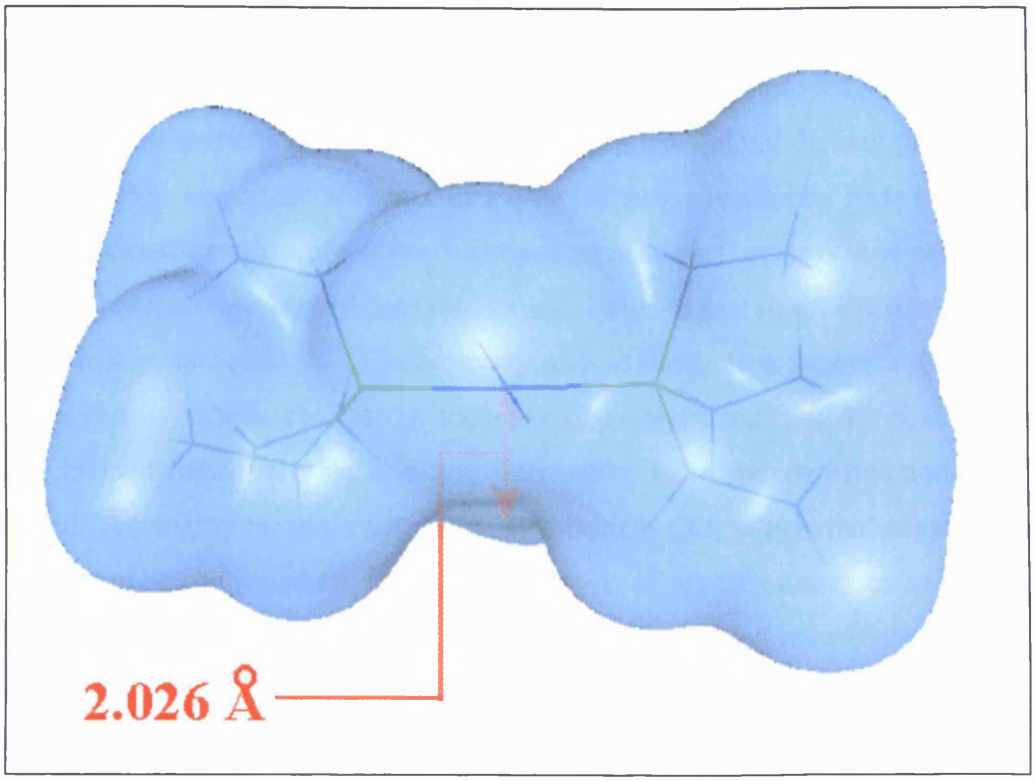
$$\ln P = -1.1862 - 0.0085\ TSA - 0.1013\ ASA_S + 0.0192\ FSA - 0.4090\ MetalSA \quad (6.3)$$

$N = 101,\ R^2 = 0.913,\ RMSE = 0.807,\ F = 251.96,\ R^2_{CV} = 0.900$

Incorporation of transition metal complexes into this model does not alter the relative significance of descriptors from eq 6.2 The largest t-ratios are still found for FSA and TSA, which act to increase and reduce ln*P*, respectively. The presence of exposed metal surface area also acts to reduce ln*P*, possibly through weak attractions between the cationic metal center and electron rich toluene molecules. The least significant descriptor, ASA_S, acts to reduce ln*P* due to the weak hydrogen bond basic properties of toluene's π-system.

In an attempt to improve on the simple assignment of 2 Å for the van der Waals radii of all metals, the electron density of the metal core at the B3LYP/Lanl2DZ level was calculated, and the 0.001 $e.au^{-3}$ isosurface was searched for the point closest to the metal nucleus. The 0.001 isosurface is widely accepted as an accurate measure of the van der Waals surface of a molecule in the gas phase.[19] In some cases, most notably the zirconium complexes Tm7 and Tm8, no such point was found within 2.5 Å, indicating that the entire metal atom is engulfed by its ligands. In most other cases, however, reasonable estimates of vdw radii were obtained, ranging from 2.026 Å for Ni in Tm2 to 2.453 Å for Fe in Tm6 (see Figure 6.1 for an example of such a calculation). Unfortunately, these calculated radii did not improve the quality of prediction of ln*P*, and in several cases made the predictions substantially worse. Thus it appears that straightforward assignment of the vdw radius of each metal as 2Å is more reliable than these more elaborate estimations.

Figure 6.1: 0.001 e.au^{-3} isosurface in a model complex $[Ni(Cl)_2(PEt_3)_2]$, showing closest approach to Ni nucleus.



As well as incorporating the transition metals into the organic dataset, it is also possible to successfully model their fluorophilicity individually, though the relative scarcity of data means any conclusions based on this must be treated with caution. Again, a three parameter model using TSA, FSA, and MetalSA is very accurate, giving $R^2 = 0.94$ and RMSE = 0.62, while no improvement is gained by adding in PSA or BenSA. As with all previous models, FSA dominates the model, with TSA and MetalSA having smaller, counteracting effects. Clearly, more data are required before any firm conclusions can be made, but this method shows much promise for the prediction of fluorophilicity of metal complexes.

6.2 Solubility in Supercritical Carbon dioxide

6.2.1 Introduction

Another green solvent that has been widely accepted and employed within industrial chemistry are supercritical fluids such as CO_2. These have many distinct properties that make them highly important solvents in many industrial reactions. Examples of such processes are decaffeination of coffee,[20] extraction of hops,[21] spices[22] and seed oils.[23,24] Supercritical solvents are highly tuneable so can be used to extract thermally sensitive material at low temperature.[25,26] The recovery of a solute from a supercritical solvent is easily achieved by simple changes in the operating conditions. Supercritical CO_2 is of particular interest industrially as it is non-flammable, inert, inexpensive, non-toxic and unregulated. The highly tuneable nature of supercritical CO_2 allows the solvating power of the solvent to be controlled *i.e.* the solvating power of the CO_2 increases as its pressure is raised. Due to these properties supercritical CO_2 has been used industrially as a solvent for sensitive extracting/separating reactions where purity of the extracted material must be of a very high standard.

There have been many studies reported into the prediction of solubility. Famini *et al*[27] used a theoretical linear solvation energy relationship (TLSER) to predict the solubility of 22 aromatic compounds in supercritical CO_2 at 14 MPa at 308K and 20 MPa at 308K. The TLSER method of Famini and Wilson[28] is a computational implementation of Kamlet and Taft's LSER approach. LSER models solvation properties as linear combinations of size, polarity, and hydrogen bonding terms. Famini *et al*'s study of solubility in supercritical CO_2 produced the following equation.

$$\log S_{CO2} = -6.037\pi_i + 10.440\varepsilon_\beta - 22.098q^- + 24.350q^+ - 8.370 \qquad (6.4)$$

$N = 19$, $R^2 = 0.928$, $Sd = 0.477$

Where π_i is a polarity/dipolarity index, ε_β is the molecular orbital hydrogen bond basicity, q^- and q^+ are the electrostatic hydrogen bond basicity and acidity respectively. While these statistics prove that the model is capable of predicting supercritical CO_2 solubility for a small set of similar molecules at the same temperature and pressure it does not fully test the ability of quantitative structure activity relationships (QSAR) predictive powers. Famini's TLSER descriptors are calculated only for the most negative and positive formal charge in the

molecule and therefore are only applicable to 1:1 complexes. This approach to calculation of the descriptors will cause the model to fail when there is multiple complexation.

Politzer *et al* [29] initially used two computational parameters based on electrostatic potentials to find correlations when comparing the solubility of naphthalene and eight indoles in four super critical fluids (C_2H_6, C_2H_4, CO_2 and CHF_3.) They then further defined three solute molecular properties:[30] (a) surface area, (b) the sum of the variance between the positive and negative electrostatic potentials of the surface and (c) a balance term parameter, which indicates the extent to which a solute's positive and negative regions can interact. These three terms were used to model the solubility of 22 aromatic molecules in various supercritical fluids with impressive accuracy. Bush *et al*[31] created a model of solubility in supercritical CO_2 based on the LFER developed by Kamlet *et al*[32]. The model was based on a dataset of 35 molecules at a constant temperature and pressure of 308 K and 28.9 MPa. A model was produced with an average error of 65 %.

A similar study has also been performed by Dongjin.[33] Here the LFER approach of Abraham[10] was used to verify the proposed retention mechanism when performing supercritical fluid chromatography with CO_2, when an organic modifier is used with an octyldecylisilane bonded phase in a packed capillary column. In this study LFER equations were constructed for different concentrations of organic modifier. This study showed that as the concentration of organic modifier increased the importance of a molecule size, polarizability and H bond bascicity on solubility decreased.

Although the scaled PSA and Abraham approach have proven reliable in modelling many solvent systems they do not allow for large changes in the density of the solvent. The highly tuneable nature of supercritical CO_2 and the range of temperatures and pressures within available data can be accounted for by the addition of a sixth descriptor, π^1, as suggested by Lagalante.[34] π^1 describes the polar/polarisability of the solvent at a specific density. It also gives a measure of the solvent's ability to induce dipolar phenomena in the solute, and is analogous to the S term in LFER, except it describes the chemical properties of the solvent and not the solute. For the Abraham descriptors only S requires an analogous solvent term as the other descriptor values for supercritical CO_2 have been

proven constant over the gas/liquid density range, via measurement made using UV visible solvatochromatic methods.[35]

Solubility values (used here as $\log S_{CO2}$, where S is in mole fraction) for molecules in CO_2 can be measured using numerous techniques.[36] There are four main categories into which all of the most commonly used techniques fall, these are i) Flow or Dynamic ii) Static iii) Chromatographic iv) Spectroscopic. The easiest and most common method for obtaining $\log S_{CO2}$ values is via the flow dynamic method. As flow dynamic is the easiest and most reliable the vast majority of data used in this study was obtained from literature that used this method.

6.2.2 Methods

A dataset of solubility in supercritical CO_2 for 67 molecules at varying temperature and pressure was compiled from various references,[29,37-61] the majority of which were acquired through the online solubility database.[62] This data was acquired from the references in the form of mole fraction, which was converted to $\log S_{CO2}$. The temperature range for this data was between 308K and 433K with a pressure range of between 74.3 bar and 410 bar.

Scaled PSA descriptors were calculated using the method outlaid in chapter 4.9
The π^1 descriptor was calculated according to the following relation.

$$\rho_r = \frac{\rho}{\rho_c} \tag{6.5}$$

$$\pi^1 = 1.15\rho_r - 0.98 \quad (\rho_r < 0.7) \tag{6.6}$$

$$\pi^1 = 0.173\rho_r - 0.37 \quad (\rho_r > 0.7) \tag{6.7}$$

Where ρ_r is the reduced density, ρ_c is the critical density of CO_2 and ρ is the density of supercritical CO_2. Values for CO_2 density and supercritical density were obtained from the NIST on-line Chemistry web book.[63]

6.2.3 Results

Various models were constructed using the scaled and unscaled polar surface area descriptors for the entire set of 830 log S_{CO2} values. The best model obtained utilised the descriptors TSA, π^1, BenSA, ASA_S and BSA_S. The model gave a disappointing R^2 value of 0.33 and an RMSE of 1.19.

A regression of the Abraham descriptors against the same dataset of 830 $logS_{CO2}$ values gave the following LFER equation. Analysis of the t-ratio for each individual coefficients indicated that the *s* coefficient of the S descriptor was insignificant, while all other descriptors are >99 % significant. For this reason the S descriptor is omitted from the model.

$$logS_{CO2} = 1.062 \ -1.778E \ -0.702A \ -0.634B \ -0.343V + 4.184\pi^1 \qquad (6.8)$$

$N = 830$, $R^2 = 0.718$, RMSE = 0.780, $R^2_{cv} = 0.713$

Analysis of these model showed three large outliers whose predicted $logS_{CO2}$ values were more than 2 log units from that of the observed $logS_{CO2}$ values. These outliers were 2,4, D, beta-carotene and piroxicam, although piroxicam is only an outlier in the LFER model and is modelled adequately using the scaled PSA method. The presence of these molecules clearly affects the accuracy and validity of both models. With the removal of the appropriate outliers from the regressions the following equations were given.

Scaled PSA

$$LogS_{CO2} = 1.109 - 0.001 \ TSA + 0.051 \ ASA_S - 0.022 \ BSA_S - 0.017 \ BenSA + 3.20 \ \pi^1 \qquad (6.9)$$

$N = 791$, $R^2 = 0.447$, RMSE = 1.050, $R^2_{cv} = 0.441$

Abraham LFER

$$logS_{CO2} = 1.124 \ -1.753E \ -0.642A \ -1.054B \ -0.22V + 4.32\pi^1 \qquad (6.10)$$

$N = 782$, $R^2 = 0.782$, RMSE = 0.659, $R^2_{cv} = 0.779$

Both models show improvement upon the removal of outliers, the greatest improvement is seen in the scaled polar surface area method with R^2 rising to 0.447 and RMSE

dropping to 1.05. Although the improvement is most notable within the scaled PSA model, the LFER of Abraham is still clearly the better of the two methods.

Although the statistics of both of these models are inferior to those offered by the supercritical CO_2 model of Famini[27] (R^2 =0.928, SD = 0.477), our models cover a much wider range of $logS_{CO2}$ values (-7.347 to -0.762 log units), a spread of over 6.5 log units. Our dataset also contains a wider range of molecules, many of which are multifunctional; our methods are also capable of predicting $logS_{CO2}$ values at varying temperature and pressure.

A list of the outliers removed from equations 6.9 and 6.10 are given in Table 6.2. The calculated descriptor values for these outliers were checked by comparing partition coefficient (logP) values calculated using the generated descriptors against observed logP values for a number of different solvent systems. Observed logP values were obtained from the MedChem2000 database[64] the results are shown in Table 2. LogP values for β-carotene were not available from the MedChem database and could not therefore be analysed in this manner.

Table 6.2: Observed and calculated logP values for outliers

Solvent	Piroxicam					2,4-D				
		Abraham LFER		Scaled PSA			Abraham LFER		Scaled PSA	
	Obs [a]	Calc	Error	Calc[b]	Error	Obs [a]	Calc	Error	Calc[b]	Error
Octanol	1.795	-0.647	2.442	1.929	-0.134	2.729	2.656	0.073	2.67	0.059
Hexadecane	-1.52	-6.178	4.658	-	-	-	-	-	-	-
PGDP	-0.07	2.752	-2.822	-	-	-	-	-	-	-
$CHCl_3$	-	-	-	-	-	1.2	2.301	-1.101	0.936	0.264
Air	-	-	-	-	-	7.75	6.499	1.251	-	-

[a] Taken from ref. [59], averaged where several data reported.

[b] Values calculated using the equation generated in 4.1.

The large difference in observed and calculated values for piroxicam using the Abraham descriptors shows clearly that the error in predicting $logS_{CO2}$ values in our model for this molecule is caused by incorrect calculation of the descriptors. Piroxicam is a zwitterionic molecule, which may account for the errors in its calculation as the fragment method used for obtaining descriptor values assumes the neutral form. For this reason piroxicam was

removed from the Abraham model. By contrast calculated Abraham descriptors give reasonable prediction of three logP values for 2,4-D. We omit this molecule from our final model since it appears there might be some problem with the experimental values used. Although we could not analyse β-carotene using the same method to determine error we still removed it from both models. One possible explanation for the erroneous predicted values of β-carotene may be due to the inability of either set of descriptors to account for the electronic effects caused by the molecule's highly conjugated sp^2 hybridized carbon backbone.

To test the importance of the π^1 descriptor and to see whether its use is warranted in these model the dataset was remodelled with the exclusion of the outliers but this time the π^1 descriptor was omitted from the model to give equations 6.11 and 6.12.

$$\log S_{CO2} = -1.954 - 0.0005TSA + 0.0546ASA_S - 0.019BSA_S - 0.015BenSA \quad (6.11)$$
$$N = 791,\ R^2 = 0.398,\ RMSE = 1.096\ R^2_{cv} = 0.389$$

$$\log S_{CO2} = -0.088 - 1.7E - 0.38A - 1.21B + 0.268S - 0.08V \quad (6.12)$$
$$N = 782,\ R^2 = 0.691,\ RMSE = 0.785\ R^2_{cv} = 0.687$$

From these statistics it can be seen clearly that the inclusion of π^1 significantly improves the quality of the results. With the removal of the π^1 descriptor the R^2 value drops by 0.5 and the RMSE increases by 0.046 for the scaled surface area method, with a similar decrease in accuracy in the Abraham LFER method.

The relative importance of each descriptor in equation 6.9 and 6.10 was assessed by analysis of the t-ratios, as listed in Table 6.3. The most significant descriptor in the scaled PSA method is BenSA, while the most significant of the Abraham descriptors is E, both of these descriptors exhibit the largest t-ratio value and a large negative coefficient, from this result it can be concluded that supercritical CO_2 is highly opposed to interacting with the substrate with a high density of π- and n-electron pairs. π^1 is seen to be highly significant in both methods, with a large positive coefficient, showing that the higher the density of the supercritical CO_2 the more the solute will be dissolved. Negative

coefficient values are also given for A and B. A larger negative t-ratio is given by B with respect to A, indicating that the stronger a substrate's hydrogen bond basicity the less it will be solvated by the supercritical CO_2. A large negative t-ratio is also given for BSA_S corroborating the conclusion drawn from the LFER that hydrogen bond bases will have less affinity for supercritical CO_2.

It should be noted that within the scaled surface area method the hydrogen bond acidity term ASA_S has a positive coefficient value within the equation. This result is in conflict with the negative A value obtained from the Abraham method. More confidence is given to the result from the Abraham LFER as the equation was proven better by statistical analysis. While the t-ratio values are in opposition it should be noted that the hydrogen bond acidity term is the least significant term in both equations and that models of almost equal quality can be made without its inclusion.

Table 6.3: Coefficients and t-ratios for equation 6.9 and 6.10

Scaled PSA				LFER			
Term	Coefficient	S.E.	t – ratio	Term	Coefficient	S.E.	t – ratio
Intercept	-1.167	0.142	-8.197	Intercept	1.128	0.108	10.428
TSA	-0.001	0.000	-4.841	V	-0.22	0.024	-8.975
ASA_S	0.048	0.011	4.151	A	-0.642	0.097	-6.628
BSA_S	-0.020	0.001	-17.308	B	-1.054	0.068	-15.576
π^1	3.161	0.377	8.389	π^1	4.323	0.239	18.071
BenSa	-0.017	0.001	-19.775	E	-1.753	0.04	-44.148

A small negative coefficient is also displayed by the size terms TSA and V, revealing that for supercritical CO_2 the exoergic dispersion forces are dominated by the endoergic cavity term. From this it can be concluded that solubility in supercritical CO_2 is favoured if the volume/size of the substrate is small. As has already been stated S was omitted from the equation, indicating that the polarity/polarisability of the substrate has no effect on its solubility. This is to be expected, as CO_2 has no permanent dipole for polar interaction and isn't very polarisable.

The negative coefficients for E, A, B, BenSA and BSA_S can be rationalised by comparing the LFER equation given in this study to that of the LFER equation for the prediction of melting point given by Platts/Saunders.[65] The LFER for melting point shows large positive coefficients for these descriptors, where a positive coefficient value indicates an increase in solid-state stability. Most of the molecules used to construct eq (6.10) are solids in the conditions used, such that the stabilising solid-state interactions must be disrupted before solvation in CO_2 can occur. While this is not equivalent to melting, the similarity of the two equations strongly suggests that solid-state effects are an important factor in determining solubility of solids in supercritical CO_2.

The two equations can be quantitatively compared by treating them as vectors and calculating the angle in between them:[66] in this case, we calculate an angle of 137° between the LFER's for solubility in CO_2 and melting point. Thus there is clearly some relationship between these two processes, although there are also substantial differences between them also. The main difference in the angle between vectors for these equations is due to the difference in *a* coefficient values. In contrast to equation 3 the LFER equation for melting point shows a distinctively higher value for A than B. The contrast between these two equations implies that supercritical CO_2 has some H-bond basicity.

We can compare the properties we have proposed for the solubility in supercritical CO_2 to those proposed by Famini.[27] Absolute comparisons cannot be drawn between our equations and those of Famini as their descriptors are calculated using a different method, and also our model includes the sixth descriptor π^1 to account for the effects of temperature and pressure. Famini's electrostatic basicity term, analogous to our BSA_S and B, also has a large negative coefficient. Famini's equation shows a negative value for the dipolar/polarisability term analogous to our S term. The value of the coefficients for Famini's polar descriptor is much less than that of the electrostatic basicity. Famini reported a positive electrostatic acidity term, suggesting an increase in H bonding acidity in the substrate would increase its solubility, again implying that supercritical CO_2 has hydrogen bond basic properties.

Famini stated that these values are due to CO_2 being harder 'non polarisable' rather than soft 'polarisable' and that harder solutes are more soluble than soft solutes. This is

consistent with both Pearson[67] and Drago[68] models of acidity and basicity. They also show the significance of the harder terms electrostatic acidity and basicity to be far higher than those of soft terms such as polarisability. The descriptors for H-bond acidity and basicity, which we have employed, are composite of both hard and soft terms, unlike those of Famini, and we therefore cannot conclude anything about hardness and softness. However, the non-significance of the S term shows that polarity of a species is insignificant to its solubility in supercritical CO_2, consistent with the conclusions of Famini.

6.3 Critical Micelle Concentration

6.3.1 Introduction

A molecule's critical micelle concentration CMC is defined as the concentration range at which individual isolated surfactant molecules begin to aggregate to form micelles due to surface activity. After the CMC is exceeded any additional surfactant added to the solution will form micelles. Once the CMC of a surfactant has been reached, many important physicochemical properties such as surface tension, conductivity, and detergency change dramatically.[69] These properties are important to many industrial and biological systems, so the ability to predict CMC directly from molecular structure is of great interest.

The relationship between molecular structure and CMC has been well documented. A typical surfactant molecule can be broken down to two components that contribute towards CMC, namely the hydrophobic (tail) and hydrophilic region (head). As the size of the hydrophobic region is increased it becomes more thermodynamically favourable for the hydrophobic regions of the surfactant molecule to minimize contact with the aqueous solution, seen as a decrease in CMC. In contrast as the size and hydrophilic properties of the head group are increased CMC rises.[69]

Linear relationships between the logCMC (typically measured mol/L) and the number of carbon atoms in a surfactant's hydrophobic tail have been defined for homologous series of linear alkyl hexaethoxylates by Rosen[70] and octaethoxylates by Merguro.[71] Ravey[72] showed a linear relationship between the number of ethylene oxide units and logCMC for dodecyl polythoxylates. Beecher[73] used both the number of carbon atoms and number of ethylene oxide units to predict CMC for a series linear alkyl ethoxylate surfactants. The following equation was produced:

$$\log CMC = A + Bm + Cn \qquad (6.13)$$

Where m and n are the number of carbon atoms and ethylene oxide units respectively A, B and C are regression co-efficient. The predictive ability of this relationship was

improved by Ravey[72] who introduced a non-linear descriptor; a cross term defined as the number of carbon atoms multiplied by number of ethylene oxide units.

There has also been great deal of success in predicting CMC using quantitative structure property relationships (QSPR). Wang *et al*[74] derived the following equation for a set of 29 linear alkyl ethoxylates and ten alkyl phenyl polyethylene oxides.

$$\log CMC = 1.930 - 0.7846KH0 - 8.871 \times 10^{-5} E_T + 0.04938D \qquad (6.14)$$

$N = 39,\ R^2 = 0.995$

Where KH0 is the Kier&Hall index of 0 order[75], E_T is the total molecule energy (in eV) and D is the dipole moment (in Debye) of the surfactant and N is the number of molecules that were used in the regression. Direct comparison of equation 2 and those proposed by Ravey[72] and Beecher revealed that equation 2 was as accurate as the previous models but had the benefit that it could be used to predict CMC not only for alkyl ethoxylates but also alkyl phenyl polyethylene oxides.

Huibers *et al*[76] used the program CODESSA[77] (Comprehensive Descriptors for Structural and Statistical Analysis) to predict CMC for a series of 77 non-ionic surfactants. The CODESSA program uses a heuristic approach to select the most appropriate descriptors from a large pool of several hundred descriptors. The study produced the following equation.

$$\log CMC = -1.802 - 0.567\ c\_KH0 + 1.054\ c\_AIC2 + 0.751\ RNNO \qquad (6.15)$$

$N = 77,\ R^2 = 0.983$

Where AIC2 is the information content index,[78] RNNO (relative number of nitrogen and oxygen) is the number of oxygen and nitrogen atoms divided by the total number of atoms in the molecule. The prefix c_ indicates that the descriptor only refers to the hydrophobic regions of the surfactant.

Huibers *et al* followed up this study by using CODESSA to derive an equation for the prediction of CMC for anionic surfactants.[79] This equation was based on a dataset of 119 sulphonates and sulphate molecules. CODESSA produced the following equation.

$$\log CMC = 1.89 - 0.314t\text{-}sum\text{-}KH0 - 0.034TDIP - 1.45h\text{-}sum\text{-}RNC \quad (6.16)$$

$N = 119$, $R^2 = 0.940$

Where t-sum-KH0 is the Kier Hall molecular connectivity indices of zeroth order[75] for all hydrophobic regions, TDIP is the total dipole of the molecule, and h-sum-RNC is the sum of the relative number of carbon atoms for hydrophilic regions.

The R^2 values for equation 6.14, 6.15 and 6.16 show that the models are of a high quality. However these models are all constructed from datasets with low diversity of functional groups, for instance many of molecules within these datasets are homologous series. The aim of this study is to try to establish a more general model for the prediction of CMC for more structurally diverse molecules such as drug molecules.

While the heuristic approach of programs such as CODESSA may find correlations that could otherwise have been missed, the models produced often forgo the clarity and interpretability of models produced using other QSPR methods. A further aim for this study is that from the models produced further information can be inferred about the physiochemical factors influencing the formation of micelles.

Although the scaled PSA method was developed for transfer processes involving two or more solutions liquid or solid phases, we hypothesise that these methods will be flexible enough to model CMC. An analogy can be made between CMC and a two pseudo-phase partition processes except the partition is between solvated and associated solute. CMC is determined by factors such as the self-association properties of water and the relative hydrophobicity of the tail, properties that can be accounted for using descriptors such as molecular size and volume. CMC is also determined by the self-association and repulsive properties of the surfactant head groups, which can be accounted for with descriptors that encompass the molecules hydrogen bonding abilities.

6.3.2 Method

Three separate datasets were compiled, the first two from previous studies of CMC by Huibers *et al.*[76,79] Dataset one contained 77 non-ionic surfactants in aqueous solution at 25°C with logCMC values ranging from -6.523 to -0.009 log units. Dataset two contained 119 anionic surfactants in aqueous solution at 40°C, 50 of these values were recorded at 25°C and their values at 40°C were calculated using the recommended ratio 1.088 and 1.030 for sulphonates and sulphates respectively, a ratio that has been established to be approximately constant for the CMC of these molecules.[79] For dataset two logCMC values ranged from -4.899 to -0.496 log units.

A third dataset was compiled from Schrier's paper,[80] this dataset contains 32 drug molecules in aqueous solution at 30°C. These molecules include analgesics, anaesthetics and antibiotics, whose logCMC values range from -6.22 to -0.60 log units. It should be noted that all of these molecules have been seen to form micelles and do not associate in a manner in which aggregate size increases continuously with increasing concentration.

The first two datasets allow direct comparison between the methods used in this study and previously published ones. Also, dataset one and the majority of dataset two can be combined, allowing our methods to be applied simultaneously to the calculation of CMC for ionic and non ionic surfactants. Dataset three was selected as it contains many drug molecules that are already of great commercial interest, and also the number of different functional groups present is significantly broader than that of any previous study of CMC.

Descriptors were generated using the method stated in 4.9. An analogous Abraham model was also constructed for each model, this allowed verification of any conclusions drawn from the PSA method. For the Abraham LFER model descriptors were calculated using the group contribution method of Platts.

It should be noted that the group contribution of Platts *et al* does not contain fragments for the anionic oxygen of the sulphonates and sulphate in dataset two, so the SMILES were changed so the anionic oxygen was treated as an oxygen in S=O. (This can be justified as every molecule in the dataset contains one anionic oxygen and hence can be regarded as

constant within regression analysis). It should also be noted that there is no scaling factor for the anionic oxygen present in dataset two; solutions to this problem are discussed later.

6.3.3 Results

6.3.3.1 Dataset 1: Non Ionic

Regression of the CMC values of dataset 1 against scaled surface area descriptors gave the following equation.

$$\log CMC = 1.282 - 0.017TSA - 0.256ASA_S + 0.067BSA_S - 0.007HalSA - 0.001BenSA \quad (6.17)$$

$N = 77,\ R^2 = 0.903,\ RMSE = 0.434,\ R^2_{cv} = 0.880,\ F = 131.5$

When the same regression is performed using PSA_U descriptor along with TSA a significantly poorer model is produced with R^2 dropping by 0.65 and the RMSE rising by 0.736 log units. Table 6.4 contains the results of statistical analysis for all models. The drop in predictive accuracy is easily clarified when the t-ratios of the descriptors in 6.17 are analysed. The t-ratios show that ASA_S and BSA_S have equal but opposing effects on CMC, i.e. as ASA_S is increased the value of logCMC predicted by equation 6.17 will decreases whereas raising BSA_S serves to increase values of logCMC calculated by equation 6.17. Hence the amalgamation of ASA_S and BSA_S to form PSA_S will create a descriptor that cannot correctly account for the physiochemical properties of CMC, as determined by equation (6.17).

Table 6.4: Statistical analysis of CMC models

Model	Descriptors	N	R^2	RMSE	R^2cv	F-ratio
Dataset 1.	TSA PSA_S		0.250	1.179	0.174	12.2
Non-ionic	TSA ASA_S BSA_S HalSA BenSA	77	0.903	0.433	0.879	131.6
	Abraham LFER		0.856	0.527	0.805	84.6
Dataset 2.	TSA PSA_S		0.851	0.338	0.839	335.3
Anionic	TSA ASA_S BSA_S [b] HalSA BenSA	119	0.871	0.320	0.855	151.9
	Abraham LFER		0.868	0.324	0.846	148.0
Combined data	TSA PSA_S		0.390	0.998	0.350	39.9
from datasets	Scaled TSA ASA_S BSA_S [a] HalSA BenSA	127	0.757	0.757	0.741	75.3
1 and 2.	Scaled TSA ASA_S BSA_S [b] HalSA BenSA O-SA		0.826	0.543	0.810	114.8
	Scaled TSA ASA_S BSA_S [b] HalSA BenSA Indicator		0.815	0.570	0.800	94.9
Dataset 3.	TSA PSA_S		0.734	0.691	0.422	66.5
Drug molecules	Scaled TSA ASA_S HalSA BenSA	32	0.909	0.418	0.829	67.7
	Abraham LFER		0.909	0.420	0.080	67.2

[a] O^- surface area include with value scaled to one

[b] O^- surface area not included

A regression of the same dataset against the Abraham descriptors produced the following equation.

$$\log CMC = 1.2066 - 1.400E + 4.06S - 3.480A + 1.522B - 3.148V \qquad (6.18)$$

$N = 77,\ R^2 = 0.856,\ RMSE = 0.527,\ R^2_{cv} = 0.805,\ F = 84.6$

The statistics of this regression are similar to, but slightly lower than, those of equation 6.17 with a slight decrease in R^2 and a slight increase in RMSE. While the statistics of equation 6.17 and 6.18 show both methods produce accurate models, the predictive power of these equations is rather less than that published by Huibers *et al* for the same dataset, eq 6.15. The main source for this loss of accuracy is that many molecules in the dataset contain large quantities of intramolecular hydrogen bonding, which has been seen to 'tie up' both acid and base atoms and thus alter their acid and base properties. Although the scaled surface area method contains fragments that account for intramolecular H-bonding around aromatic rings, the definitions are not comprehensive enough to accurately calculate the properties of molecules such as sucrose monooleate and beta-dodecyl

maltoside, which are seen to be the two largest outliers for the scaled surface method (residuals of -1.037 and 0.967 respectively).

6.3.3.2 Dataset 2: Ionic Surfactants

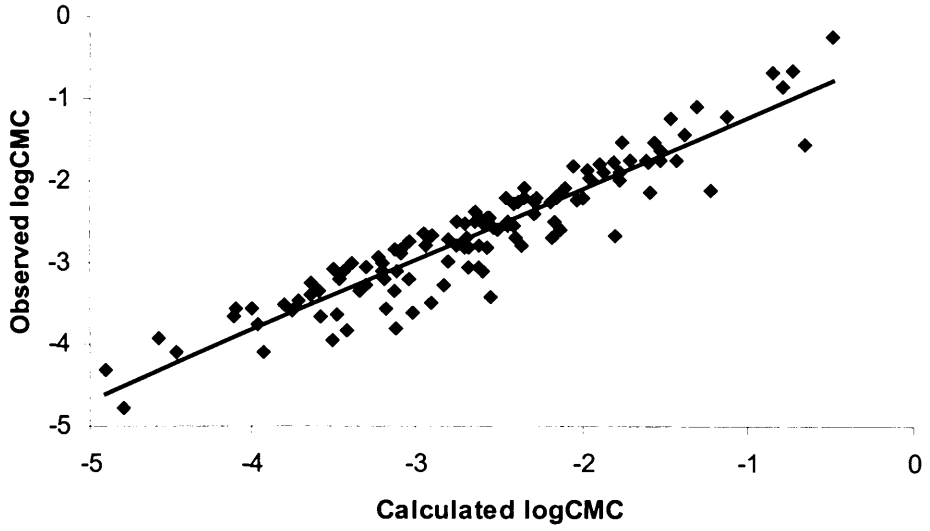
Table 6.4 contains various statistical analysis of the models produced from the 119 ionic surfactants of dataset two. As there are no defined experimental values for the anionic oxygen of the sulphonates and sulphates in the Abraham scales of A and B the following measures were taken to account for this in the scaled surface area method.

1. Models were created in which the O^- surface area was incorporated into the definition of BSA_S and scaled with a value of one. This made it approximately equivalent to oxygen in sulphoxide.
2. The O^- surface area was removed from the definition of BSA_S and allocated its own descriptor, which was termed O^-SA.
3. The O^- surface area was completely omitted from the BSA_S descriptor.

The results showed that the three methods stated above make very little difference to the statistics of the model with R^2 values being 0.866, 0.875 and 0.871 for methods 1, 2, and 3 respectively. This is because the surface areas of O^- are almost constant through the dataset with values only ranging from 16.2 to 20.9 $Å^2$, and a standard deviation of 1.79 $Å^2$. Not only are the surface areas of the O^- constant but also their occurrence, with one present for each surfactant in the dataset.

In this case, the models made containing only PSA_S and TSA are only marginally worse than those made with the five parameter scaled surface area descriptors. This major change is due to the fact that 97% of molecules in dataset 1 have some H-bond acidity, whereas in dataset 2 only 16 % of the surfactants contain H-bond acidic groups. Hence for dataset two PSA_S is dominated by BSA_S, and PSA_S and BSA_S are almost interchangeable. While the overall statistics of the models change very little, fourteen of the nineteen surfactants that do possess H-bond acidity show a marked improvement when ASA_S and BSA_S are used separately. Figure 6.2 shows the correlation between observed vs. calculated logCMC values for the five parameter scaled surface area model.

Figure 6.2: Observed logCMC values vs. calculated logCMC for ionic dataset. (dataset 2)



The range of different functional groups that fall into the defined fragments of our scaling factors is very narrow, with only 12 different types occurring (2 acid fragments and 10 base). The lack of structural diversity in the surfactants accounts for the fact that when the scaling factors for ASA_S and BSA_S are removed, the model produced is statistically comparable to that of the model that includes scaling factors.

The model produced by Huibers *et al*[79] for this dataset gave an R^2 value of 0.94. The cause for the loss of accuracy here is due to the occurrence of numerous series of surfactants in which the tail group remains constant and the position of the headgroup moves from the terminal to the medial position along the carbon chain e.g. 1-dodecanesulphate through to 6-dodecanesulphate. Using the LFER relationship approach of Abraham and the group contribution method of Platts, the descriptors for these surfactants will be calculated to be equal. Although the surface area method is 3D, the changes in descriptors for these series are so subtle that the associated changes in CMC are not modelled fully. Additional accuracy could be achieved by including topological descriptors such as KH0, but this would negate the physical interpretability of such a model and prevent comparison with other solvation phenomena.

Although the results of these models show that molecular surface areas can predict CMC with reasonable accuracy, the nature of the dataset, with few surfactants displaying any

hydrogen bond acidic properties and the highly similar structure of the molecules, does not challenge the PSA descriptor enough to merit its splitting into ASA and BSA or its scaling.

6.3.3.3 Combined dataset

One of the strengths of the molecular surface area models is that it is possible to easily combine data from dataset one and two and simultaneously predict logCMC for anionic and non-ionic surfactants. The two datasets could not be combined directly as dataset one is measured at 25°C and dataset two is measured at 40°C, and the temperature dependence of CMC is well documented.[81] 50 surfactants from dataset 2 had CMC values recorded at 25°C which could be combined with the 77 surfactants of dataset one which were also recorded at 25°C. It is not possible to back extrapolate the CMC of remaining 69 surfactants in dataset two using the ratio stated earlier, as a number of the structures have Kraft points higher than 25°C, meaning that micelles would not be formed at 25°C and that calculated values would be physically meaningless. It not possible to combine the surfactants of dataset three as their CMC values were observed at 30°C, and no single ratio can be assigned to such a diverse set.

MLRA was performed against this combined dataset of 127 CMC and their molecular surface area descriptors, the statistical analysis for these models is shown in Table 6.4. The models that employ only the PSA_U and TSA descriptors are clearly incapable of modelling CMC. Small improvements result from separating PSA_U, but acceptable statistics only result when the ASA_U and BSA_U are scaled to account for their H-bonding strengths, yielding an increase of *ca.* 25 % in R^2 and a drop of 0.13 in RMSE over unscaled models.

If the anionic oxygen surface area is removed from BSA_S and included as a separated descriptor O^-SA further improvement is seen to the model, causing the R^2 and R^2_{cv} to increase by 7% and RMSE to decrease by 0.21. Within this combined dataset, the presence of the O^-SA descriptor is not so constant as it is in dataset two, hence its inclusion has higher significance within this model. Given the fairly constant values of O^- SA, a model of similar quality can be using an indicator variable, defined as one for

anionic and zero for non-ionic surfactants, in place of O^-SA (R^2 = 0.815, RMSE = 0.57). Although these six-parameter models are less accurate than separate models, the ability to simultaneously model CMC for charged and uncharged surfactants is a unique feature of this method.

In order to establish the predictive capability of this method, 25% of the data points were randomly removed to create a test set while the remaining 75% were remodelled; the equation generated from this regression used to predict the CMC values of the test set. This process was repeated a further three times to include all molecules in at least one test set. The results show that the models are capable of accurate prediction, with R^2 = 0.818 and RMSE = 0.550 when averaged overall four test sets.

6.3.3.4 Dataset 3: structurally diverse drug molecules

Dataset three represents the most challenging of all three datasets as it contains a wider range of functional groups and molecular structures than any other model of CMC. Analysis of ASA_S and BSA_S (and A and B) for this dataset showed that the two descriptors correlate with an accuracy of about 86%. This high correlation means that if both descriptors were to be included in the same model errors would be generated and interpretability of the model and predicted values would be unreliable. It should be noted that for all previous models low correlations between ASA_S and BSA_S (and A and B) were found. Stepwise multiple linear regressions of the 5 scaled surface area descriptors revealed that BSA_S was insignificant and highly accurate models could be made without the inclusion of BSA_S. Similar conclusions were reached for B in LFER models. Thus, BSA_S and B are omitted from all reported models.

The results in Table 6.4 reveal again that PSA_S and TSA alone cannot model CMC as well as split four-parameter models. Scaling of the ASA descriptor yields only a 3% increase in R^2 and a 0.067 decrease in RMSE, but a notable increase of 45% is seen in R^2_{cv}. The statistics of the LFER model are almost identical to the 4 parameter scaled surface area model except in R^2_{cv} where a difference of 75 % is reported. The difference in R^2_{cv} between the LFER and four parameter scaled surface area model is due entirely to the inability of the LFER method to predict the value of Actinomycin D when it is

omitted during the cross-validation procedure, whereas the four-parameter surface area model predicts the CMC value of Actinomycin D with reasonable accuracy. It should be noted that the structural diversity of this dataset is much higher that that of the previous sets, with 32 of our 46 defined fragments being employed in the assignment of scaling factors.

6.3.4 Discussion

The significance of each descriptor in a model is given by its t-ratio, rather than its coefficient, so these are given in table 6.5. The pattern in t-ratios is fairly constant across models, with the molecular size descriptors giving large negative values for all four datasets, indicating that larger molecules will form micelles at lower concentrations. This relationship is well established and has been stated in previous studies[81] of CMC e.g. CMC decreases by half for every methylene added to the chain for ionic surfactants.

Table 6.5.A:. t-ratios for best surface area models of CMC

	Ionic	Non-ionic	Ionic/nonionic	Structurally diverse
Intercept	6.36	5.79	2.51	2.56
TSA_S	-25.97	-22.87	-19.84	-4.45
ASA_S	-3.70	-19.07	-14.63	-7.18
BSA_S	6.18	20.04	16.13	N/A
HalSA	-2.83	-11.66	-8.64	-2.35
BenSA	1.26	-0.24	0.62	-3.21
O-SA	-2.04	N/A	0.03	N/A

Table 6.5.B:. t-ratios for Abraham LFER models of CMC

	Ionic	Non-ionic	Structurally diverse
Intercept	2.85	4.36	1.65
E	-1.31	-4.73	-5.45
S	2.54	7.84	4.42
A	-3.65	-8.52	-7.28
B	1.66	4.86	N/A
V	-24.92	-19.31	-1.55

The molecular size term are the most significant in all models except for dataset three. The drop in significance of the size terms for dataset three is due to the complex 3D structures of the surfactants. The surfactants in dataset one and two can be predominantly split into their hydrophobic tail and hydrophilic head components, with the hydrophobic regions being mainly straight hydrocarbon chains. These can intertwine easily during micelle formation due to their flexibility, making the process of surfactant-surfactant interaction on micellization fairly constant over all surfactants. This simple intertwining is not possible for many of the molecules in dataset three such as Thioridazine and Actinomycin D. Thus the size terms for dataset three is forced to account for both the enthalpic and entropic factors that are needed to create a cavity in the solvent and for the surfactants and the self-association properties upon micelle formation.

Hydrogen bond acidity (A and ASA_S) and Basicity (B and BSA_S) descriptors are also fairly constant in their t-ratio values throughout, with hydrogen bond acidity terms giving negative values and hydrogen bond basicity terms giving positive values. This result indicates that stronger hydrogen bonding acidic surfactants will form micelles at lower concentrations than weaker hydrogen bond acidic surfactants, while increasing the hydrogen bond basicity of a surfactant acts to raise its CMC.

Further insight into the role of the hydrogen bonding descriptors can be gained by comparing coefficient values from the LFER logCMC models to those for Abraham's model of aqueous solubility ($logS_w$).[82] This comparison allows us to infer how proportionately the descriptors are representing the ability of the surfactant to interact with water and interact with themselves during aggregation. The LFER for log S_w gives large positive co-efficents for A and B of 0.65 and 3.39 respectively. The LFER models of CMC also show positive values for B indicating that surfactants with a larger hydrogen bond basicity can interact with water favourably thus reducing their ability to form micelles and raising CMC.

The hydrogen bond acidity descriptor gives small positive coefficients in the $logS_w$ model but a large negative value in models of CMC. The difference in these coefficient values indicates that A is predominantly describing self-association effects of the surfactants and not surfactant water-interactions. It is not surprising that of the two descriptors, A and B, it is A that contains the information for self-association. Using the definitions of H-Bond

acidity and basicity stated in this study it is possible for a molecule to be only a hydrogen bond base, i.e. contain no hydrogen attached to oxygen and nitrogen, whereas it is impossible for a surfactant to display only H-bond acidic properties. Hence any molecule with hydrogen bond acid groups will also contain hydrogen bond basic groups, giving rise to strong self-association interactions. The surface area descriptor BenSA is not highly significant in any of the models of CMC, nor is its value constant throughout all systems. For datasets one and three negative values of BenSA are displayed, as expected since it is known that the addition of one phenyl group is roughly equivalent in its effects on CMC as three methylene groups. The positive value of BenSA for dataset two is perhaps due to the lack of structural diversity here, since all phenyl rings are attached to an electron withdrawing SO_3^- group.

HalSA's t-ratios are relatively small and negative throughout all models of CMC. This negative value can be easily attributed the fact that halogens are hydrophobic in nature. The E and S descriptors of the LFER approach are again easily interpreted by comparison to the LFER for log S_w. The coefficient for the S descriptor is positive in both the CMC models and logS_w indicating that more polar surfactants will associate more preferentially with water and thus raise CMC. The coefficient for E is negative in both equations implying that surfactants with a high density of π- and n-electron pairs would rather interact with each other than with water, presumably through dispersion forces.

6.4 Conclusions

6.4.1 Conclusions: fluorophilicity

A model for the prediction of fluorophilicity of 93 organic molecules has been constructed. Statistical analysis has shown this model to be as accurate as those produced from previous studies, with $R^2 = 0.94$ and RMS error = 0.64. The use of exposed surface areas meant it was possible to include molecules omitted from previous studies, such as three silicon containing compounds. More importantly, given the interest in catalysis in fluorous phases, it was also possible to include eight transition metal complexes bearing fluorous ligands along with the organic molecules. This yielded a good correlation of observed *vs.* calculated values, with $R^2 = 0.91$ and RMS = 0.807. These models employed total, polar, fluorous and metal surface areas. Initially, we employed van der Waals radii of 2Å for all transition metals. Subsequent attempts were made to calculate more accurate radii by finding the distance of the 0.001 $e.au^{-3}$ isosurface from the metal nucleus. While the radii generated from this method appeared physically realistic, the resultant model was marginally inferior to that of its predecessor.

6.4.2 Conclusions: Supercritical CO_2

We have developed a model for the prediction of solubility in supercritical CO_2 based on 65 molecules and 781 data points. The model had a temperature range from 308 K to 435 K and a pressure range of 74.3 bar to 410 bar.

The equation given by our scaled PSA model has a very large positive coefficient value for the π^1 descriptor. The equation also showed a relatively large negative coefficient values for the hydrogen bond basicity descriptors, B and BSA_S, the polar/polarisabilty term S and the aromatic carbon surface area BenSA. From these coefficient values it was determined that solubility in supercritical CO_2 is increased if the density of the CO_2 is increased. The solubility of a species in CO_2 will be favoured if its π- electron density and H-bond basicity is low. To a lesser extent the solubility in supercritical CO_2 is favoured if the molecule is small. These conclusions concur with those stated by other methods.

6.4.3 Conclusions: CMC

Models have been developed for the prediction of CMC for anionic and non-ionic surfactants, using the LFER approach of Abraham and surface area method of Saunders *et al.* The models produced from these datasets were slightly less accurate, but are more generally applicable, than those produced from previous studies. Furthermore, they allow detailed physical analysis, giving an insight into the factors determining CMC. This analysis was possible as the same descriptors were used throughout and not chosen for each set from a larger pool of descriptors. Thus a model was created that combined molecules from the ionic and non-ionic datasets using a modified version of the surface area approach. A reasonable correlation of observed vs. calculated CMC values were seen for this model, though statistics are slightly worse than for separate models of neutral and ionic surfactants. The predictive capability of this model was confirmed by the construction of training and test sets, which showed that logCMC could be predicted to 0.55 log units.

As the structural diversity included in these surfactant models was very narrow, a model was made that included drug molecules which included a wide range of functional groups and molecular structures. The best model produced for this set was the scaled surface area model, which had an R^2 value of 0.91 and an RMSE of 0.42. This model was also examined to give information about the micellization process the findings were in keeping with those of other models and more importantly are physically valid.

6.5 Future work

Alternate route to the calculation of vdw radii for metal complexes could lead to more accurate values of MetalSA. Also alternative routes toward the calculation of MetalSA would be beneficial as the current quantum mechanical methods may be a little to demanding for large libraries of molecules.

6.6 References

1. I.T. Horvath, *Accounts Chem. Res.*, **1998**. 31. 641.
2. E. deWolf, G. vanKoten, J. Deelman, *Chem. Soc. Rev.*, **1999**. 28. 37.
3. L.P. Barthel-Rosa, J.A. Gladysz, *Coord. Chem. Rev.*, **1999**. 192. 587.
4. E.G. Hope, A.M. Stuart, *J. Fluorine Chem.*, **1999**. 75. 100.
5. I.T. Horvath, J. Rabai, *Science*, **1994**. 266. 72.
6. L.E. Kiss, J. Rabai, L. Varga, I. Kovesdi, *Synlett*, **1998**. 1243.
7. C. Rocaboy, D. Rutherford, B.L. Bennett, J.A. Gladysz, *J. Phys. Org. Chem.*, **2000**. 13. 596.
8. L.E. Kiss, I. Kovesdi, J. Rabai, *J. Fluor. Chem.*, **2001**. 108. 95.
9. F.T.T. Huque, K. Jones, R.A. Saunders, J.A. Platts, *J. Fluorine Chem.*, **2002**. 119. 115.
10. M.H. Abraham, *Chem. Soc. Rev.*, **1993**. 22. 73.
11. J.A. Platts, D. Butina, M.H. Abraham, A. Hersey, *J. Chem. Inf. Comput. Sci.*, **1999**. 39. 835.
12. P.R. Duchowicz, F.M. Fernandez, E.A. Castro., *J. Fluorine. Chem.*, **2004**. 43. 125.
13. E. dewolf, P. Ruelle, J. vanden-Brocke, B. Deelman, G. van-Koten, *J. Phys. Chem*, **2004**. 108. 1458.
14. J.A. Gladysz, *http://www.chemie.uni-erlangen.de/gladysz/research/partition.html*, **2004**.
15. M. Svensson, S. Humbel, R.D.J. Froese, T. Matsubara, S. Sieber, K. Morokuma, *J. Phys. Chem.*, **1996**. 100. 19357.
16. M.J. Frisch, *Gausian03*, **2003**.
17. V. Herrera, P.J.F. de Rege, I.T. Horvath, T. Le Husebo, R.P. Hughes, *Inorg. Chem. Commun.*, **1998**. 1. 197.
18. M.A. Guillevic, C. Rocaboy, A.M. Arif, I.T. Horvath, J.A. Gladysz, *Organometallics*, **1998**. 17. 707.
19. R.F.W. Bader, M.T. Carroll, J.R. Cheeseman, C. Chang., *J. Amer. Chem. Soc.*, **1987**. 109. 7968.
20. K. Zosel, *Angew. Chem. Int. Ed. Engl.*, **1978**. 17. 702.
21. D. Laws, N. Bath, J. Pickett, *J. Inst. Brew.*, **1977**. 86. 39.
22. P. Hurbert, O. Vitxthumb, *Angew. Chem. Int. Ed. Engl.*, **1978**. 17. 710.
23. J.P. Friedrich, G.R. List, A.J. Heakin, *J. Am. Oil Chem. Soc.*, **1982**. 59. 288.
24. E. Stahl, E. Shultz, H. Mangold, *Agric. Food. Chem.*, **1980**. 29. 1153.
25. M. McHugh, V. Krukonis, *Supercritical Fluid Extraction: Principles and Practice*. 1986, Boston: Butterworths.
26. E. Stahl, K.W. Quirin, D. Gerard, *Dense Gases for Extraction and Refining*. 1978, Berlin: Springer.
27. G.R. Famini, L.Y. Wilson, *J. Phys. Org. Chem.*, **1993**. 6. 539.
28. G.R. Famini, L.Y. Wilson, *J. Med. Chem.*, **1991**. 34. 1668.
29. P. Politzer, P. Lane, J.S. Murray, T. Brinck, *J. Phys. Chem.*, **1992**. 96. 7938.
30. J.S. Murray, P. Lane, T. Brinck, P. Politzer, *J. Phys. Chem.*, **1993**. 97. 5144.
31. D. Bush, C.A. Eckert, *Fluid Phase Equilib.*, **1998**. 151. 479.
32. M.J. Kamlet, R.W. Taft, J.L.M. Abbound, *J. Am. Chem. Soc.*, **1977**. 91. 8325.
33. P. Dongjin, L. Wenboa, L.L. Milton, J.D. Weckwerth, P.W. Carr, *J. Chromatogr.*, **1996**. 753. 291.
34. A.F. Lagalante, T.J. Bruno, *J. Phys. Chem. B*, **1998**. 102. 907.
35. R.D. Smith, S.L. Frye, C.R. Yonker, R.W. Gale, *J. Phys. Chem.*, **1987**. 91. 3059.

36. L. Guo-tang, N. Kunio, *J. Supercrit. Fluids.*, **1996**. 9. 152.
37. I. Ashour, H. Hammam, *J. Supercrit. Fluids*, **1993**. 6. 3.
38. S.L.J. Yun, K.K. Liong, G.S. Gurdial, N.R. Foster, *Ind. Eng. Chem. Res.*, **1991**. 30. 2476.
39. S.S.T. Ting, S.J. Macnaughton, D.L. Tomasko, N.R. Foster, *Ind. Eng. Chem. Res.*, **1993**. 32. 1471.
40. A. Laitinen, M. Jantti, *J. Chem. Eng. Data*, **1996**. 41. 1418.
41. L. Barna, J.M. Blanchard, E. Rauzy, C. Berro, *J. Chem. Eng. Data*, **1996**. 41. 1466.
42. S.J. Macnaughton, N.R. Foster, *Ind. Eng. Chem. Res.*, **1994**. 33. 2757.
43. S.J. Macnaughton, I. Kikic, N.R. Foster, P. Alessi, A. Cortesi, I. Colombo, *J. Chem. Eng. Data*, **1996**. 41. 1083.
44. S.J. Macnaughton, I. Kikic, G. Rovedo, N.R. Foster, P. Alessi, *J. Chem. Eng. Data*, **1995**. 40. 593.
45. D.J. Miller, S.B. Hawthorne, A.A. Clifford, S. Zhu, *J. Chem. Eng. Data*, **1996**. 41. 779.
46. M. Richter, H. Sovova, *Fluid Phase Equilib.*, **1993**. 85. 285.
47. S.N. Joung, K.P. Yoo, *J. Chem. Eng. Data*, **1998**. 43. 9.
48. J.W. Hampson, *J. Chem. Eng. Data*, **1996**. 41. 97.
49. V.S. Smith, A.S. Teja, *J. Chem. Eng. Data*, **1996**. 41. 923.
50. M. Johannsen, G. Brunner, *J. Chem. Eng. Data*, **1997**. 42. 106.
51. H. Uchiyama, K. Mishima, S. Oka, M. Ezawa, M. Ide, T. Takai, P.W. Park, *J. Chem. Eng. Data*, **1997**. 42. 570.
52. M. AshrafKhorassani, M.T. Combs, L.T. Taylor, F.K. Schweighardt, P.S. Mathias, *J. Chem. Eng. Data*, **1997**. 42. 636.
53. M.S. Curren, R.C. Burk, *J. Chem. Eng. Data*, **1997**. 42. 727.
54. P. Alessi, A. Cortesi, I. Kikic, N.R. Foster, S.J. Macnaughton, I. Colombo, *Ind. Eng. Chem. Res.*, **1996**. 35. 4718.
55. M. Mukhopadhyay, P. Srinivas, *Ind. Eng. Chem. Res.*, **1996**. 35. 4713.
56. T. Yonemoto, T. Charoensombutamon, R. Kobayashi, *Fluid Phase Equilib.*, **1990**. 55. 217.
57. M.L. Cygnarowicz, R.J. Maxwell, W.D. Seider, *Fluid Phase Equilib.*, **1990**. 59. 57.
58. S. Li, G.S. Varadarajan, S. Hartland, *Fluid Phase Equilib.*, **1991**. 68. 263.
59. M. Johannsen, G. Brunner, *Fluid Phase Equilib.*, **1994**. 95. 215.
60. J.W. Chen, F.N. Tsai, *Fluid Phase Equilib.*, **1995**. 107. 189.
61. V.J. Krukonis, R.T. Kurnik, *J. Chem. Eng. Data*, **1985**. 30. 247.
62. R. Gupta, D. Varlamov, *Online Supercritical Solubility Database. (http://database.iem.ac.ru/scf/general.html)*, **2000**.
63. N. WEB-BOOK, *NIST Standard Reference Database Number 69 - February 2000 Release. Online Reference book (http://webbook.nist.gov/chemistry/)*.
64. A.J.Leo, *MedChem software, Med. Chem. Biobyte*, **1997**.
65. R.A. Saunders, J.A. Platts, *Unpublished Work*, **2002**.
66. Y. Ishihama, N. Asakawa, *J. Pharm. Sci.*, **1999**. 88. 1305.
67. R.G. Pearson, *J. Am. Chem. Soc.*, **1963**. 85. 3533.
68. G.C. Vogel, S.J. Drago, *J. Am. Chem. Soc.*, **1971**. 93. 6014.
69. M.J. Rosen, *Surfactants and interfacial phenomena*. 1989, New York: Wiley,. 110.
70. M.J. Rosen, *J. Colloid Interface Sci.*, **1976**. 56. 320.

71. K. Meguro, Y. Takasawa, N. Kawahashi, Y. Tabata, M. Ueno, *J. Colloid Interface Sci.*, **1981**. 83. 50.
72. J.C. Ravey, A. Fherbi, M.J. Stebe, *J. Prog. Colloid Polym. Sci.*, **1988**. 76. 234.
73. P. Beecher, *Dispers. Sci. Technol.*, **1984**. 5. 81.
74. Z.W. Wang, G.Z. Li, X.Y. Zhang, R.K. Wang, A.J. Lou, *Colloid Surf. A-Physicochem. Eng. Asp.*, **2002**. 197. 37.
75. L.B. Kier, L.M. Hall, *Molecular connectivity in chemistry and Drug Research*. 1986, New York: Academic Press.
76. P.D.T. Huibers, V.S. Lobanov, A.R. Katritzky, D.O. Shah, M. Karelson, *Langmuir*, **1996**. 12. 1462.
77. A.R. Katritzky, V.S. Lobanov, M. Karelson, *Chem. Soc. Rev.*, **1995**. 24. 279.
78. M.I. Stankevich, I.V. Stankevich, N.S. Zefirov, *Russian chemical Reviews*, **1988**. 57. 191.
79. P.D.T. Huibers, V.S. Lobanov, A.R. Katritzky, D.O. Shah, M. Karelson, *J. Colloid Interface Sci.*, **1997**. 187. 113.
80. S. Schreier, S.V.P. Malheiros, E. de Paula, *Biochim. Biophys. Acta-Biomembr.*, **2000**. 1508. 210.
81. D. Attwood, A.T. Florence, *Surfactant systems Their chemistry, pharmacy and biology*. 1983, London, New York,: Chapman and Hall.
82. M.H. Abraham, J. Le, *J. Pharm. Sci.*, **1999**. 88. 868.

7.1. Closing Remarks

For the models of $\log P_{oct}$, $\log P_{CHCl}$ and $\log P_{cyc}$ produced in chapter 4 it was seen that our new surface area method was comparable to that of the well established LFER method of Abraham[1] with descriptors calculated using the group contribution method of Platts.[2] It was seen from the t-ratios that the coefficient values of these equations were physically meaningful and in agreement with other studies.

While these models showed that for organic molecules there was little difference between the two methods the true strength of the new surface area method was seen in the modelling of inorganic molecules. While there are a numerous QSAR and QSPR methods for the calculation of organic molecules there are no well-established methods for the prediction of inorganic molecules. The ability of the new surface area method to model the properties of inorganic molecules was verified by the modelling of two commercially important properties, these were the partitioning of platinum containing drugs between octanol/water and chloroform/water, and the fluorophilicity of a number of important transition metal catalysis used in fluorous biphasic catalysis.

For the partition models of the platinum drugs it was seen that through the assignment of appropriate scaling factors for polar atoms attached to the central platinum atom that organic and inorganic data could be modelled simultaneously. The ability of our methods to combine organic and inorganic data is particularly useful in the field of QSAR as there is a large amount of data available for organic molecules but very little for inorganic molecules, so the ability to combine this data means that models can be made with larger datasets and a greater degree of confidence can be placed in the models.

Through the modelling of a number of important biological properties it was also seen that the new surface area descriptors were capable of modelling important biological data such as uptake of gases by plants and cellular uptake, while these models were created for only organic molecules it can be proposed that if data were available for inorganic molecules these could be modelled and values predicted. This is an area of particular interest industrially as the resistance of the chemotherapy drug cis platin has been attributed to reduced cellular uptake.

The ability of our new surface area method to model a wide range of physiochemical properties has been verified through the successful modelling of a wide range of diverse properties including aqueous/organic partitions, organic/organic partitions, solubility of gases, solubility of solids and CMC.

From the model of CMC it was seen that through the application of an anionic surface area descriptor, it was possible to model the properties of anionic molecules either independently or in combination with uncharged molecules. Again this simultaneous modelling of two different types of molecule offers substantial improvement of other QSAR methods as the size of datasets can be increased and the confidence in the models can be raised.

7.2 References

1. M.H. Abraham, *Chem. Soc. Revs.*, **1993**. 22. 73.
2. J.A. Platts, D. Butina, M.H. Abraham, A. Hersey, *J. Chem. Inf. Comput. Sci.*, **1999**. 39. 835.